2 avril 2017

TP 1 - Statistique descriptive, Analyse en composantes principales

UV : **SY09**
Branche : **Génie Informatique**
Filière : **Fouille de Données et Décisionnel**
Auteurs : **LU Han - HAMONNAIS Raphaël**

Table des matières

1. Statistique descriptive

1.1 Notes de SY02 au semestre de printemps 2016

1.1.1 Analyse descriptive

Description des données

Le jeu de données contient des informations relatives aux 296 étudiants inscrits à l'UV SY02 au semestre de printemps 2016 ainsi que les résultats qu'ils ont obtenus. On a donc 296 mesures (individus de la population) représentés par onze variables : 8 qualitatives et 3 quantitatives.
Liste des variables qualitatives :

» **nom** : nom de l'étudiant, au format texte (noms anonymisés, au format « *Etu1, Etu2, ...* »).

» **specialite** : spécialité de l'étudiant, $\in$ {GB, GI, GM, GP, GSM, GSU, HuTech, ISS, TC}.

» **niveau** : le numéro du semestre actuel de l'étudiant, $\in \{1, 2, 3, 4, 5, 6\}$.

» **statut** : vaut $\{UTC, Echange\}$ selon que c'est un étudiant de l'UTC ou bien un étudiant originaire d'une autre université effectuant une partie de ses études à l'UTC.

» **dernier.diplome.obtenu** : le dernier diplôme obtenu par l'étudiant, $\in$ {AUTRE 1ER CYCLE, AUTRE 2E CYCLE, AUTRE DIPLOME SUPERIEUR, BAC, BTS, CPGE, DEUG, DUT, ETRANGER SECONDAIRE, ETRANGER SUPERIEUR, INGENIEUR, LICENCE, NA'S}.

» **correcteur.median** et **correcteur.final** : le correcteur ayant corrigé la copie (nons anonymisés au format « *Cor 1, ...* »).

» **resultat** Le résultat de l'étudiant, de A à F ou ABS s'il a été absent à l'un des deux examens (final ou médian). F et Fx signifient que l'étudiant n'a pas obtenu l'UV, les autres notes qu'il l'a obtenue.

Liste des variables quantitatives :

» **note.median** : la note obtenue au médian ($\in [0; 20]$).

» **note.final** : la note obtenue au final ($\in [0; 20]$).

» **note.totale** : la moyenne des deux notes, pondérée par l'importance de chaque note.

Données manquantes

On remarque qu'il manque certaines informations dans les données (R le spécifie avec le mot clé « *NA's* » pour *Not Available* en anglais) :

» Correcteur du médian et/ou du final : élève absent au médian et/ou au final.

» Dernier diplôme obtenu : donnée manquante pour les étudiants en échange.

- Note au médian et/ou au final : donnée manquante pour les étudiants absents ou qui ont abandonné.
- Résultat : donnée « manquante » pour R qui considère la valeur *ABS* comme étant manquant car *resultat* est une variable qualitative ordonnée et *ABS* ne fait pas partie des différents niveau d'ordre. Représente donc des élèves absents qui n'ont pas obtenu l'UV.

Intuitions statistiques : liens supposés entre les variables

De manière logique, on pourrait penser que les notes entre le médian et le final sont liées, un élève bon au médian étant plus à même de réussir le final et inversement. De même, la formation d'origine de l'étudiant devrait fortement influencer l'obtention de l'UV. Un étudiant venant de tronc commun (diplôme d'origine BAC) a logiquement plus de change de réussir qu'un élève venant de DUT. Le correcteur ne devrait logiquement pas influencer les notes, la notion d'équité entre les copies étant très importante.

1.1.2 Etude des liens statistique entre les variables

Pré-requis

On considère que les données ont été préalablement nettoyées en enlevant tous les étudiants absents, c'est à dire qui n'ont pas eu l'UV pour cause d'absence accidentelle ou d'abandon pur et simple.

Procédure

Nous utilisons le test d'indépendance du χ^2 qui permet de vérifier l'indépendance de deux variables X et Y :

- Hypothèse nulle H_0 : les deux variables X et Y sont indépendantes.
- Hypothèse H_1 : les deux variables X et Y ne sont pas indépendantes.
- On rejette l'hypothèse nulle lorsque *p-value* est inférieure ou égale à 0,05. La valeur *p-value* représente la probabilité d'obtenir la même valeur (ou une valeur encore plus extrême) du test si l'hypothèse nulle était vraie.
- Toutes les cases du tableau de contingence doivent avoir une valeur supérieure ou égale à 5. Le tableau de contingence est un tableau à double entrée représentant les effectifs partiels des observations en fonction des variables X en ligne et Y en colonne.

Lien statistique entre le résultat et le diplôme d'origine des étudiants

La première étape fut de nettoyer les données en supprimant les 6 étudiants dont le statut vaut *Echange* car leur diplôme d'origine n'est pas renseigné.
Nous avons ensuite créé le tableau de contingence, qui donne pour chaque valeur de la variable *resultat* le nombre d'élèves ayant obtenu ce résultat en fonction de leur diplôme d'origine : on obtient un effectif pour chaque couple *diplôme d'origine/résultat*.

	F	Fx	E	D	C	B	A
AUTRE 1ER CYCLE	1	0	2	1	2	0	0
AUTRE 2E CYCLE	1	0	0	0	0	0	0
AUTRE DIPLOME SUPERIEUR	0	1	0	1	1	0	0
BAC	4	11	12	16	26	24	13
BTS	2	1	0	1	1	2	0
CPGE	10	6	6	6	5	4	2
DEUG	1	0	0	1	1	0	0
DUT	20	12	15	17	15	8	2
ETRANGER SECONDAIRE	1	0	1	0	1	0	1
ETRANGER SUPERIEUR	3	2	0	0	4	1	1
INGENIEUR	0	0	0	1	0	0	0
LICENCE	2	0	1	0	2	3	1

(a) Tableau de contingence entre résultat et diplôme d'origine

AUTRE 1ER CYCLE	6
AUTRE 2E CYCLE	1
AUTRE DIPLOME SUPERIEUR	3
BAC	106
BTS	7
CPGE	39
DEUG	3
DUT	89
ETRANGER SECONDAIRE	4
ETRANGER SUPERIEUR	11
INGENIEUR	1
LICENCE	9

(b) Effectif total par diplôme

FIGURE 1.1 – Effectifs de la population par diplôme d'origine et tableau de contingence entre les variables *résultat* et *diplôme d'origine*.

On remarque immédiatement que les conditions du test ne sont pas respectées : il a y une majorité des effectifs qui sont inférieurs à 5 dans le tableau de contingence (Figure 1.1a). Et c'est tout à fait normal si l'on regarde les effectifs d'étudiants (Figure 1.1b) en fonction de leur diplôme d'origine dans la population totale : seuls les diplômes BAC (étudiant venant de tronc commun), DUT et CPGE sont convenablement représentés. On peut tout de même effectuer le test du χ^2 tout en sachant que les conditions ne sont pas respectées, afin de vérifier l'importance de ces conditions. On obtient alors une *p-value* égale à 0.2572. On conserve donc l'hypothèse H_0 d'indépendance des variables, tout en sachant que le résultat et probablement faux.

Correction des effectifs Afin de respecter les conditions du test de χ^2, nous n'avons gardé que les lignes correspondant aux diplômes d'origines de type *BAC*, *DUT* et *CPGE*. Il a aussi fallut regrouper les deux premières et deux dernières classes de *résultat* en sommant les effectifs des élèves ayant obtenu *F* ou *Fx* et ceux ayant eu *A* ou *B*.

On obtient ainsi un nouveau tableau de contingence :

	F-Fx	E	D	C	B-A
BAC	15	12	16	26	37
CPGE	16	6	6	5	6
DUT	32	15	17	15	10

FIGURE 1.2 – Tableau de contingence **corrigé** entre résultat et diplôme d'origine.

Les conditions du test sont alors réunies et on obtient une *p-value* égale à 0.000288. On rejette donc l'hypothèse H_0 d'indépendance avec confiance : le diplôme d'origine d'un étudiant a un impact significatif sur son résultat final à l'UV de statistiques SY02.

Voici une représentation graphique de la fréquence des résultats en fonction du diplôme d'origine :

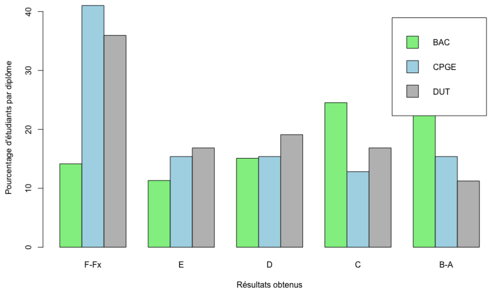


FIGURE 1.3 – Fréquence par diplôme d'origine et résultat

On remarque que 34% des élèves ayant comme dernier diplôme le BAC ont obtenu A ou B tandis que seulement 14% d'entre-eux n'ont pas réussi l'UV. On remarque aussi que la courbe est croissante : plus le résultat est bon, plus la fréquence d'obtention augmente. C'est l'inverse pour les élèves provenant de DUT ou CPGE : moins le résultat est bon, plus la fréquence d'élèves obtenant ce résultat augmente (environ 40% des élèves en provenance d'un DUT ou de CPGE n'ont pas obtenu l'UV).

Lien statistique entre le résultat et la spécialité des étudiants

La spécialité des étudiants correspond à leur cursus actuel : GB, GI, GM, GP, GSM, GSU, HuTech, ISS, ou TC. En s'intéressant à l'effectif de chaque classe (Figure 1.4), on remarque de suite qu'il va falloir en considérer quelques unes seulement pour respecter les conditions du test.

GB	GI	GM	GP	GSM	GSU	HuTech	ISS	TC
65	45	61	16	53	40	1	4	11

FIGURE 1.4 – Effectif par spécialité

Les classes telles que *HuTech*, *ISS*, *TC* voire même *GP* ne sont pas assez représentées pour être en mesure de respecter les conditions du tests d'indépendance du χ^2.

Voici les deux tableaux de contingence obtenus, le premier sur toutes les données (Figure 1.5a), le second (Figure 1.5b) en sélectionnant certaines classes de la variable *spécialité* puis en regroupant les effectifs de certaines classes de la variable *résultat*.

	F	Fx	E	D	C	B	A
GB	8	8	8	8	13	11	5
GI	7	5	4	14	7	4	2
GM	9	8	5	8	14	9	5
GP	5	1	2	2	5	1	0
GSM	14	6	8	8	8	6	2
GSU	3	6	6	4	9	10	2
HuTech	1	0	0	0	0	0	0
ISS	2	0	0	0	1	1	0
TC	0	0	4	0	1	0	4

(a) Tableau de contingence entre résultat et spécialité

	F-Fx	E-D	C	B-A
GB	16	16	13	16
GI	12	18	7	6
GM	17	13	14	14
GSM	20	16	8	8
GSU	9	10	9	12

(b) Tableau de contingence **corrigé** entre résultat et spécialité

FIGURE 1.5 – Tableaux de contingence entre résultats et spécialité.

Les deux tests du χ^2 effectués sur le tableau 1.5a et sur le tableau 1.5b donnent respectivement les résultats suivants :

- Tableau de contingence d'origine
  p-value = 0.02087781.
  On réfute (à tort) l'hypothèse H_0 d'indépendance des variables *résultat* et *spécialité*.
- Tableau de contingence corrigé
  p-value = 0.4673856.
  On accepte l'hypothèse H_0 d'indépendance des variables *résultat* et *spécialité* avec confiance.

Conclusion : la spécialité d'un étudiant n'a pas d'impact sur son résultat.

Lien statistique entre le résultat et le niveau des étudiants

Le niveau des étudiants correspond à leur semestre actuel de branche, de 1 à 6 (deux semestres par an sur un cursus de trois ans).

Voici les deux tableaux de contingences obtenus, le premier (Figure 1.6a) sans regroupement de classes et le second (Figure 1.6b) en regroupant les classes *F-Fx-E* et *C-B* et en ne considérant pas les étudiants de niveau 3,5 et 6 (effectif trop peu important).

	F	Fx	E	D	C	B	A
1	0	3	5	8	10	9	6
2	26	16	18	26	40	21	7
3	0	2	1	0	2	0	0
4	19	11	8	10	4	11	7
5	1	0	4	0	0	0	0
6	3	2	1	0	2	1	0

(a) Tableau de contingence entre résultat et niveau

	F-Fx-E	D	C-B	A
1	8	8	19	6
2	60	26	61	7
4	38	10	15	7

(b) Tableau de contingence **corrigé** entre résultat et niveau

FIGURE 1.6 – Tableaux de contingence entre résultats et niveau.

Les deux tests du χ^2 effectués sur le tableau 1.6a et sur le tableau 1.6b donnent respectivement les résultats suivants :

- Tableau de contingence d'origine
  p-value = 0.0003517831.
  On réfute l'hypothèse H_0 d'indépendance des variables *résultat* et *niveau*.
- Tableau de contingence corrigé
  p-value = 0.003523739.
  On réfute l'hypothèse H_0 d'indépendance des variables *résultat* et *niveau*.

Conclusion : le niveau d'un étudiant impacte son résultat. On observe par exemple dans la figure 1.7 que plus le niveau de l'étudiant augmente, moins il a de chance d'obtenir l'UV. 42% de *F* ou *Fx* pour les étudiants de niveau 4 contre 7% seulement pour ceux de niveau 1. A l'opposé, ces derniers sont 80% à obtenir l'UV avec un résultat entre *D* et *A* contre 45% pour les étudiants de niveau 4. Les étudiants de niveau 2 se situent au milieu, avec 60% pour un résultat compris entre *D* et *A* et 27% de non obtention de l'UV.

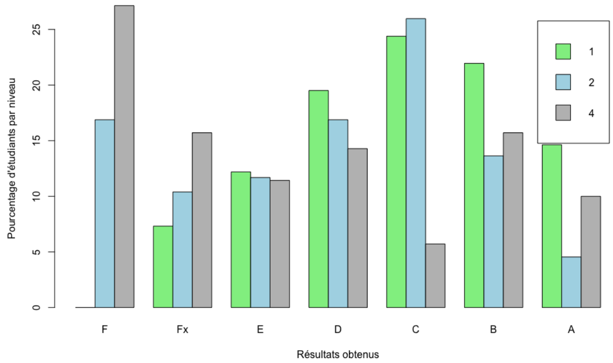


FIGURE 1.7 – Fréquence d'obtention d'un résultat en fonction du niveau de l'étudiant

Lien statistique entre le résultat et les correcteurs

Selon toute vraisemblance, le correcteur ne devrait pas influer sur le résultat d'un étudiant à l'UV. Nous allons tout de même nous en assurer. Le principe est le même que précédemment : construire le tableau de contingence, regrouper des classes au besoin pour respecter les conditions du test du χ^2 et effectuer ce test afin de déterminer si les variables sont indépendantes ou non.

Voici le premier tableau de contingence obtenu entre les variable *notes médian* et *correcteur médian* :

	0	1.5	3	3.5	4	4.5	5	5.5	6	6.5	7	7.5	8	8.5	9	9.5	10	10.5	11	11.5	12	12.5	13	13.5	14	14.5	15	15.5	16	16.5	17	17.5	18	18.5	19	19.5	20
Cor1	0	1	0	0	1	0	0	2	0	0	0	1	0	1	0	2	2	2	2	0	2	0	1	0	2	1	0	2	2	0	0	0	0	0	0	0	0
Cor2	0	0	0	0	0	0	0	2	1	0	1	5	2	3	1	1	0	3	1	2	2	2	2	1	2	0	2	7	4	1	0	0	3	0	0	0	0
Cor4	0	0	0	0	1	1	2	0	2	0	2	2	1	3	6	2	3	3	1	4	5	1	1	2	3	2	0	0	1	0	0	1	0	0	0	0	0
Cor5	0	2	1	0	1	0	0	0	0	3	4	0	2	2	3	2	3	2	3	0	1	0	4	2	4	0	2	2	0	2	0	1	0	0	0	2	1
Cor6	1	0	0	0	1	1	1	1	0	4	0	0	2	1	1	0	4	3	3	4	3	1	1	4	0	1	1	1	3	1	1	1	1	1	1	1	0
Cor7	0	0	0	2	1	1	4	2	1	1	0	3	2	3	0	4	3	1	1	1	3	3	2	2	0	1	1	0	3	1	2	1	0	0	0	0	0
Cor8	0	0	0	1	2	0	1	0	3	0	0	1	0	0	2	1	1	0	2	0	0	1	1	1	1	2	0	1	0	1	1	1	0	1	0	0	0

FIGURE 1.8 – Tableau de contingence entre les notes du médian et les correcteur

Ce tableau est inutilisable pour le test d'indépendance du χ^2. Il a donc fallut regrouper les observation en classes.

Voici le second tableau de contingence obtenu après regroupement :

	0-7	8-10	11-13	14-20
Cor1	5	7	5	7
Cor2	9	10	10	19
Cor4	10	18	14	7
Cor5	11	14	10	14
Cor6	9	11	16	13
Cor7	15	13	12	9
Cor8	8	4	5	8

FIGURE 1.9 – Tableau de contingence **après regroupement** entre les notes du médian et les correcteur

Le test du χ^2 effectué sur le second tableau de contingence donne une *p-value* égale à 0.499. On conserve donc l'hypothèse $H0$ d'indépendance. Le même principe a été appliqué sur le couple de variable *note final* et *correcteur final* à la différence qu'il y a trois classes au lieu de 4 afin de respecter les conditions du test. La *p-value* obtenue vaut 0.115. On conserve ici aussi l'hypothèse $H0$ d'indépendance.

Les notes obtenues au médian et au final étant fortement corrélées au résultat obtenu (Figure 1.10), on peut conclure en affirmant que les variables *correcteur* et *résultat* sont indépendantes.

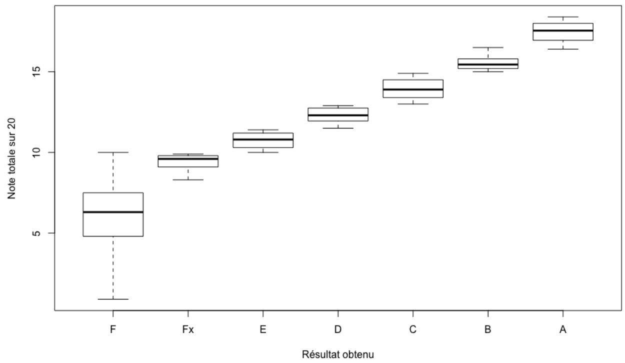


FIGURE 1.10 – Distribution des résultats obtenus en fonction de la note finale

1.2 Données Crabs

1.2.1 Analyse descriptive

Description des données

Le jeu de données présente 200 crabes classés en fonction de leur espèce et de leur sexe. Chaque individu est aussi décrit par 5 caractéristiques morphologiques quantitatives.
Liste des variables qualitatives :

- **sex** : le sexe du crabe, *M* pour mâle, *F* pour femelle.
- **sp** : espèce à laquelle appartient un individu, *O* pour *Orange*, *B* pour *Bleu*.

Liste des variables quantitatives :

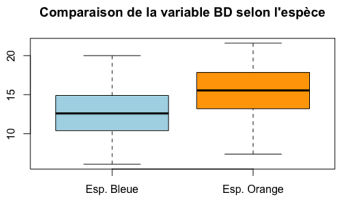
- **FL** : taille de la frontale.
- **RW** : largeur de la queue.
- **CL** : longueur de la coquille.
- **CW** : largeur de la coque.
- **BD** : la profondeur du corps.

Il existe aussi une autre variable numérique, purement utilitaire, qui est l'index :

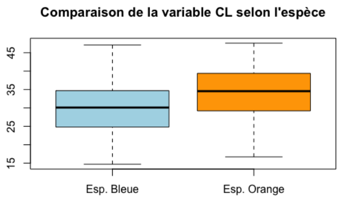
- [**1-50**] : mâles d'espèce bleue.
- [**51-100**] : femelles d'espèce bleue.
- [**101-150**] : mâles d'espèce orange.
- [**151-200**] : femelles d'espèce orange.

Données morphologiques

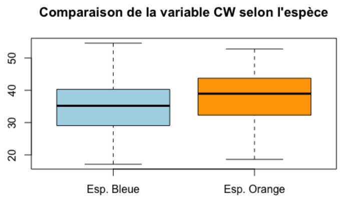
Les deux figures suivantes présentent les distributions des variables morphologiques en fonction de l'espèce (Figure 1.11) et du sexe (Figure 1.12).



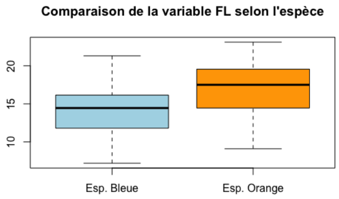
(a) Variable BD



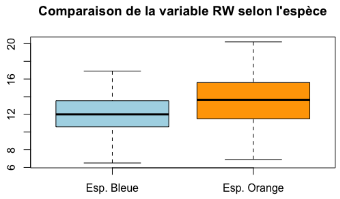
(b) Variable CL



(c) Variable CW

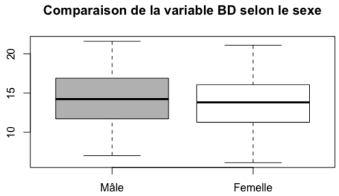


(d) Variable FL

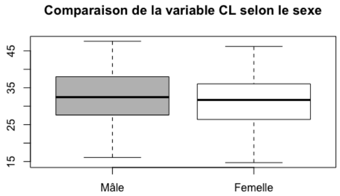


(e) Variable RW

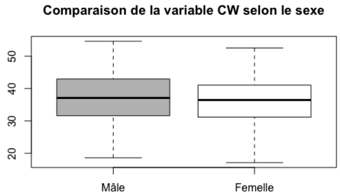
FIGURE 1.11 – Distributions des variables morphométriques en fonction de l'espèce.



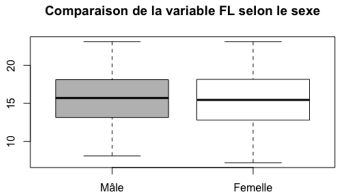
(a) Variable BD



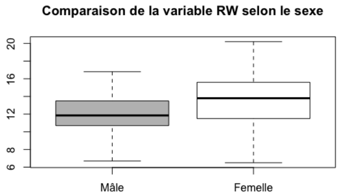
(b) Variable CL



(c) Variable CW



(d) Variable FL



(e) Variable RW

FIGURE 1.12 – Distributions des variables morphométriques en fonction du sexe.

Graphiquement, il semble difficile de séparer les individus en fonction de leur sexe ou de leur espèce. Comparer les populations de même sexe mais d'espèce différente, de même espèce mais de sexe différent ou encore de sexe et espèce différents donne des résultats similaires. Aucune de ces variables, prise individuellement, ne nous permet d'identifier le sexe ou l'espèce d'un individu.

1.2.2 Etude de la corrélation entre les variables morphologiques

La corrélation entre les cinq variables est très forte, comme le montre le tableau de corrélation suivant :

TABLE 1.1 – Corrélations entre les variables quantitatives de la population de crabes.

	FL	RW	CL	CW	BD
FL	1.00	0.91	0.98	0.96	0.99
RW	0.91	1.00	0.89	0.90	0.89
CL	0.98	0.89	1.00	1.00	0.98
CW	0.96	0.90	1.00	1.00	0.97
BD	0.99	0.89	0.98	0.97	1.00

De plus, toutes les variables sont corrélées positivement entre-elles. La raison est simple : les observations représentent des mesures morphométriques, c'est à dire des distances relevées sur les corps des crabes. Plus un crabe est grand, plus ces mesures vont grandir, et inversement. Pour s'affranchir de ce phénomène, on pourra pratiquer une analyse en composantes principales sur les données. Les nouvelles variables identifiées auront alors une très faible corrélation entre elles. Il sera alors peut-être possible de différencier le sexe et l'espèce des individus en fonction de leur mesures morphométriques analysées en fonction des nouveaux axes trouvées lors de l'ACP (voir section 2.3).

1.3 Données Pima

Les données Pima consistent en plusieurs mesures effectuée sur une population de femmes amérindiennes, le but étant d'observer les facteurs aggravant de diabète dans une population plus touchée que de moyenne par cette maladie. L'échantillon est composé 532 individus décrits par 8 variables différentes, toutes quantitatives sauf une qui détermine l'absence ou la présence d'une condition diabétique chez l'individu.

1.3.1 Analyse descriptive

Liste des variables

Variables quantitatives :

» **npreg** : nombre de grossesses.

» **glu** : taux plasmatique de glucose.

» **bp** : pression artérielle diastolique.

» **skin** : épaisseur du pli cutané au niveau du triceps.

» **bmi** : indice de masse corporelle.

» **ped** : fonction de pedigree du diabète, c'est à dire une mesure de l'influence génétique espérée des proches, affectés ou non par le diabète, sur le risque éventuel du sujet.

» **age** : l'age du sujet.

Variable qualitative :

» **z** : diabétique si $\mathbf{z} = 2$;

Domaine de définition et statistiques descriptives des variables

Table 1.2 – Résumé des variables composant le jeu de données `Pima`.

npreg	glu	bp	skin	bmi
Min. : 0.000	Min. : 56.00	Min. : 24.00	Min. : 7.00	Min. :18.20
1st Qu. : 1.000	1st Qu. : 98.75	1st Qu. : 64.00	1st Qu. :22.00	1st Qu. :27.88
Median : 2.000	Median :115.00	Median : 72.00	Median :29.00	Median :32.80
Mean : 3.517	Mean :121.03	Mean : 71.51	Mean :29.18	Mean :32.89
3rd Qu. : 5.000	3rd Qu. :141.25	3rd Qu. : 80.00	3rd Qu. :36.00	3rd Qu. :36.90
Max. :17.000	Max. :199.00	Max. :110.00	Max. :99.00	Max. :67.10

ped	age	z
Min. :0.0850	Min. :21.00	1 :355
1st Qu. :0.2587	1st Qu. :23.00	2 :177
Median :0.4160	Median :28.00	
Mean :0.5030	Mean :31.61	
3rd Qu. :0.6585	3rd Qu. :38.00	
Max. :2.4200	Max. :81.00	

La première chose importante que nous remarquons est la différence d'effectif entre les types d'individus : il y a deux fois plus d'individus non diabétiques (355) que de diabétiques (177). Le choix des représentations graphiques et tests statistiques est alors important. Travailler sur l'effectif simple avec par exemple des diagrammes en bâton donnera des résultat complètement faux. Des boîtes à moustache par contre seront plus à même de pointer des disparités entre les deux types d'individus.

Corrélation entre les variables quantitatives Cette étude de la corrélation Tableau 1.3, sans prendre en compte le facteur diabète, ne nous apprend pas grand chose sinon que les variables sont peu corrélées entre-elles, mise à part l'âge et la grossesse. Ces deux dernières sont légèrement corrélées positivement : plus la femme interrogée est âgé, plus il y a de chance qu'elle ait eu plusieurs grossesses dans sa vie.

Table 1.3 – Corrélations entre les variables quantitatives du jeu de données `Pima`.

	npreg	glu	bp	skin	bmi	ped	age
npreg	1.000	0.125	0.205	0.095	0.009	0.007	0.641
glu	0.125	1.000	0.219	0.227	0.247	0.166	0.279
bp	0.205	0.219	1.000	0.226	0.307	0.008	0.347
skin	0.095	0.227	0.226	1.000	0.647	0.119	0.161
bmi	0.009	0.247	0.307	0.647	1.000	0.151	0.073
ped	0.007	0.166	0.008	0.119	0.151	1.000	0.072
age	0.641	0.279	0.347	0.161	0.073	0.072	1.000

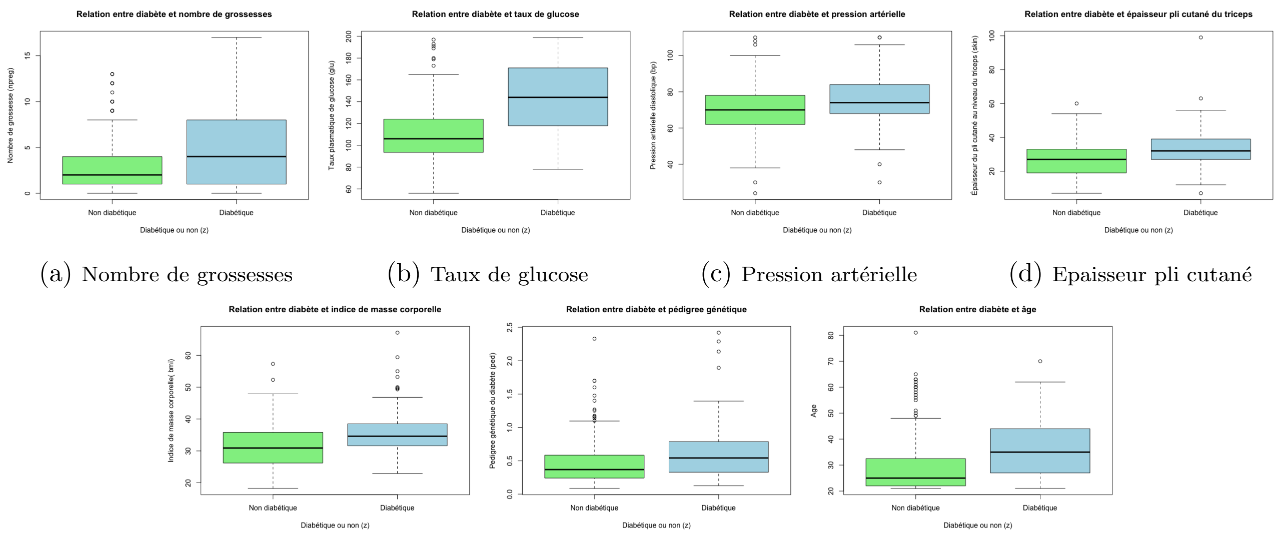
1.3.2 Etude des liens statistique entre le facteur diabète z et les autres variables

Pour étudier les liens statistiques entre le facteur diabète et les autres variables, voici la procédure que nous allons suivre :

- » Effectuer des représentations de boîte à moustache de l'ensemble des variables en fonction du facteur diabète.
- » Valider statistiquement les observations graphiques avec un test du χ^2 effectué sur le tableau de contingence vérifiant les conditions du test.
- » Représenter graphiquement le tableau de fréquence afin d'avoir une vision graphique de l'influence (ou de la non influence) du facteur diabète sur la variable étudiée. Le tableau de fréquence correspond au tableau de contingence exprimé en pourcentage de représentation d'une valeur par rapport à la population totale de la classe à laquelle elle appartient.

Note : la couleur verte représente la population non diabétique, la couleur bleu les individus souffrant de diabète.

Représentation des variables en boîte à moustache en fonction du facteur diabète



(a) Nombre de grossesses (b) Taux de glucose (c) Pression artérielle (d) Epaisseur pli cutané

(e) Indice de masse corp. (f) Pedigree génétique (g) Age

FIGURE 1.13 – Distributions des variables en fonction du diabète.

Visuellement, il semblerait que le facteur diabète est lié à l'âge, au nombre de grossesses et au taux de glucose dans le sang. Les quatre autres variables ne semblent pas diverger de façon remarquable en fonction du diabète, même si quelques différences sont déjà notables.

Tests du χ^2 d'indépendance des variables

- » **Diabète vs. Nombre de grossesses**
  p-value = 2.215e-08
  Conclusion : les variables ne sont pas indépendantes.
- » **Diabète vs. Taux plasmatique de glucose**
  p-value < 2.2e-16

Conclusion : les variables ne sont pas indépendantes, et c'est tout à fait logique, le diabète étant une maladie où le manque d'insuline — hormone en charge de la régulation du glucose — cause une élévation anormale du taux de ce dernier.

» **Diabète vs. Pression artérielle diastolique**
p-value = 0.0001862
Conclusion : les variables ne sont pas indépendantes.

» **Diabète vs. Epaisseur du pli cutané au niveau du triceps**
p-value = 9.256e-06
Conclusion : les variables ne sont pas indépendantes.

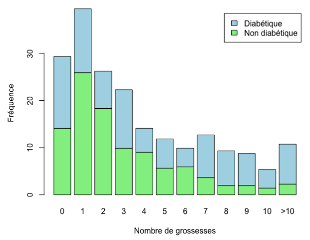
» **Diabète vs. Indice de masse corporelle**
p-value = 3.476e-11
Conclusion : les variables ne sont pas indépendantes.

» **Diabète vs. Fonction de pedigree génétique du diabète**
p-value = 1.65e-05
Conclusion : les variables ne sont pas indépendantes.

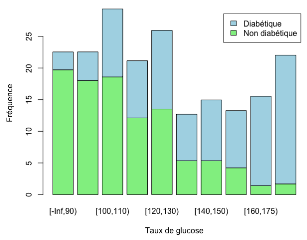
» **Diabète vs. Age**
p-value = 9.442e-14
Conclusion : les variables ne sont pas indépendantes.

Contrairement à ce qu'on pensait précédemment, on se rend compte que toutes les variables sont plus ou moins influencées par le facteur diabète. Pour certaines, comme le taux de glucose, cela se comprend facilement. Pour d'autres, l'explication est inconnue. Nous allons par la suite représenter graphiquement les tableaux de fréquence afin de comprendre l'effet du diabète sur ces variables et inversement.

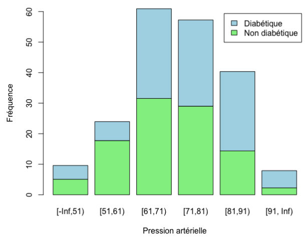
Représentation graphique des tableaux de fréquence en fonction du facteur diabète



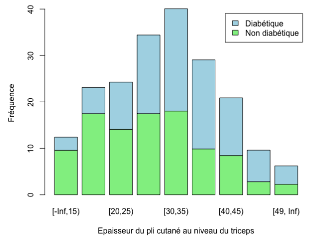
(a) Nombre de grossesses



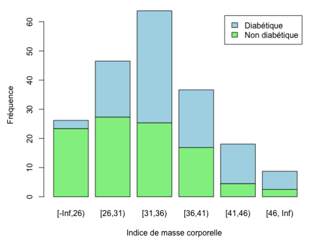
(b) Taux de glucose



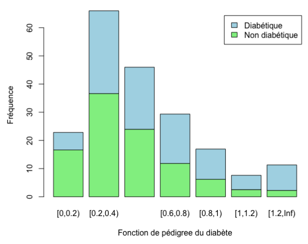
(c) Pression artérielle



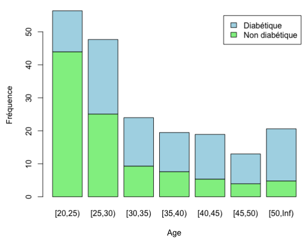
(d) Epaisseur pli cutané



(e) Indice de masse corp.



(f) Pedigree génétique



(g) Age

FIGURE 1.14 – Fréquences observées en fonction du facteur diabète.

Rappel : vert = non diabétique ; bleu = diabétique.

Ces graphiques nous donnent deux informations : la fréquence d'observation au sein d'une population diabétique ou non diabétique, mais aussi la différence de fréquence entre ces deux populations. C'est à notre sens cette deuxième information qui est la plus importante si on veut comprendre l'impact du facteur diabète sur les variables.

Analyses de l'interdépendance entre le diabète et les autres variables

- **Diabète vs. Nombre de grossesses**
  Pour un nombre de grossesses inférieur à 7, la différence de fréquence entre les deux populations n'est pas flagrante mais tends à défavoriser l'apparition de diabète. Par contre, à partir de 7 grossesses et plus, il y a une écrasante majorité d'individus atteints de diabète. Le fait d'être enceinte ou pas nous paraissant indépendant du diabète, notre conclusion est que c'est un nombre répété de grossesses qui va favoriser l'apparition du diabète (et non pas le diabète qui induit un nombre élevé de grossesses).
- **Diabète vs. Taux plasmatique de glucose**
  Comme dit précédemment, le manque d'insuline dans le corps implique un fort taux de glucose dans le sang. On peut dire que ce taux de glucose est un facteur indicatif de la présence ou de l'absence d'une condition diabétique.
- **Diabète vs. Pression artérielle diastolique**
  Ces deux facteurs sont liés, mais assez faiblement selon nous. Les individus diabétiques ont une légère tendance à avoir une pression artérielle plus élevée que les non diabétiques. Nous n'avons pas d'explication causale, à savoir est-ce le diabète qui favorise l'hypertension (et pourquoi) ou bien est-ce l'hypertension qui favorise le diabète ?
- **Diabète vs. Epaisseur du pli cutané au niveau du triceps**
  Les remarques sont les mêmes que pour la pression artérielle. On observe un effet du diabète, indiquant qu'un pli cutané plus épais est plus fréquent chez les diabétiques. Il nous semble aberrant que ce soit là la cause du diabète, aussi en déduit-on que c'est le facteur diabète qui induit cette augmentation d'épaisseur. L'explication est surement lié à l'augmentation de masse corporelle, comme on peut le voir juste après.
- **Diabète vs. Indice de masse corporelle**
  La différence de fréquence entre les deux populations diabétique et non diabétique est flagrante : un indice élevé de masse corporelle est révélateur d'une condition diabétique. Et plus cet indice est élevé, plus les chance de diabète augmentent.
- **Diabète vs. Fonction de pedigree génétique du diabète**
  Plus cette fonction de pedigree génétique du diabète augmente, plus les risques de diabète sont vérifiés.
- **Diabète vs. Age**
  Le diabète tend à toucher majoritairement les personnes adultes et plus âgées. Les individus de moins de 25 ans sont très peu à présenter un diabète.

2. Analyse en composantes principales

2.1 Exercice théorique

Soient :
$D_p = \frac{1}{n} I_n$ la matrice de poids des n individus (chaque individu possède une importance égale).
$M = I_p$ la matrice de poids des p variables (chaque variable possède une importance égale).

2.1.1 Axes factoriels de l'ACP et pourcentage d'inertie expliquée

Données de départ Les données de départ, après avoir supprimé les correcteurs 2 et 3 qui n'ont respectivement pas corrigé le final et le médian, on obtient la matrice suivante :

	moy.median	std.median	moy.final	std.final
Cor1	10.71	3.90	10.94	4.58
Cor4	10.23	3.04	13.43	4.34
Cor5	10.98	4.41	11.83	3.97
Cor6	11.50	4.30	13.41	4.88
Cor7	10.12	4.03	11.90	4.44
Cor8	10.74	4.65	11.40	4.87

Centrage en colonnes Soit X la matrice précédente centrée en colonne :

	moy.median	std.median	moy.final	std.final
Cor1	-0.01	-0.16	-1.21	0.07
Cor4	-0.48	-1.01	1.28	-0.17
Cor5	0.27	0.36	-0.32	-0.54
Cor6	0.79	0.25	1.26	0.36
Cor7	-0.59	-0.03	-0.25	-0.07
Cor8	0.03	0.59	-0.76	0.36

Le centrage en colonne s'obtient en soustrayant à chaque valeur la moyenne de la colonne correspondante : on obtient une somme de chaque colonne nulle.

Matrice de covariance On calcule la matrice V de covariance entre les variables à l'aide de la formule $V = X^T D_p X = \frac{1}{6} X^T X$ $(n = 6)$.

	moy.median	std.median	moy.final	std.final
moy.median	0.211	0.134	0.071	0.046
std.median	0.134	0.265	-0.226	0.045
moy.final	0.071	-0.226	0.908	0.013
std.final	0.046	0.045	0.013	0.099

Remarquons que la commande R `cov.wt(X, method = 'ML')` donne le même résultat.

Axes principaux d'inertie et inertie expliquée On va ici utiliser la méthode R `eigen` afin de diagonaliser la matrice de covariance. On obtient les valeurs propres suivantes :

$$\lambda_1 = 0.9799, \quad \lambda_2 = 0.3675, \quad \lambda_3 = 0.0832 \quad et \quad \lambda_4 = 0.0520$$

Les vecteurs propres associés sont :

$$u_1 = \begin{pmatrix} -0.04 \\ 0.29 \\ -0.96 \\ -0.00 \end{pmatrix} \quad u_2 = \begin{pmatrix} -0.70 \\ -0.65 \\ -0.17 \\ -0.24 \end{pmatrix} \quad u_3 = \begin{pmatrix} -0.23 \\ -0.09 \\ -0.02 \\ 0.97 \end{pmatrix} \quad u_4 = \begin{pmatrix} 0.67 \\ -0.70 \\ -0.24 \\ 0.09 \end{pmatrix}$$

Ces quatre vecteurs propres représentent les quatre axes factoriels de l'ACP définis par les quatre variables quantitatives. L'inertie expliquée par chacun des axes correspond à sa valeur propre associé. On obtient le tableau suivant :

	λ_1	λ_2	λ_3	λ_4
Inertie expliquée	0.98	0.37	0.08	0.05
Pourcentage inertie expliquée	66.10	24.79	5.61	3.51

2.1.2 Calcul composantes principales et représentation individus

Le calcul de la matrice C des composantes principales se fait avec la formule $C = XMU = XU$ avec U matrice des vecteurs propres en colonne et $M = I_p$:

	Comp.1	Comp.2	Comp.3	Comp.4
Cor1	1.11	0.30	0.11	0.40
Cor4	-1.50	0.81	0.01	0.06
Cor5	0.40	-0.23	-0.61	-0.04
Cor6	-1.16	-1.02	0.12	0.08
Cor7	0.25	0.49	0.08	-0.32
Cor8	0.90	-0.35	0.30	-0.18

Ce sont les coordonnées des observations dans le nouvel espace vectoriel dont les axes sont les vecteurs propres trouvés précédemment.

Voici la représentation des individus (correcteurs) dans le premier plan factoriel (construit à partir des deux premiers axes factoriels) :

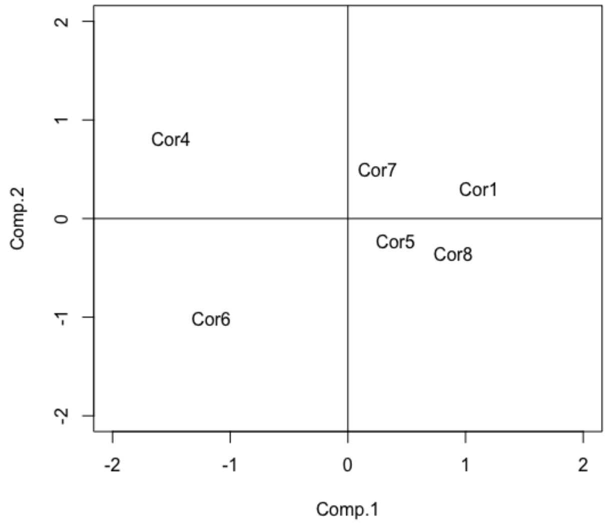


FIGURE 2.1 – Individus (correcteurs) dans le premier plan factoriel de l'ACP

2.1.3 Représentation des variables d'origine dans le premier plan factoriel

Afin de représenter les variables d'origine dans le premier plan factoriel, on va tout d'abord calculer leur corrélation. C'est à dire calculer une mesure de la représentation des variables d'origine par les nouveaux axes trouvés à l'aide de l'ACP. Si par exemple la première variable a une corrélation égale à $|1|$ avec le premier axe factoriel, alors ils représentent exactement la même chose. Si à l'inverse la corrélation vaut 0, on pourra dire que cette variable ne s'exprime absolument pas sur le premier axe factoriel.

On calcule la matrice de corrélation à l'aide de la fonction `cor` de R :

	Comp.1	Comp.2	Comp.3	Comp.4
moy.median	-0.08	-0.93	-0.15	0.33
std.median	0.57	-0.76	-0.05	-0.31
moy.final	-0.99	-0.11	-0.01	-0.06
std.final	-0.00	-0.46	0.89	0.06

On reporte maintenant ces mesures sur le premier plan factoriel afin de représenter les variables d'origine sur ce dernier :

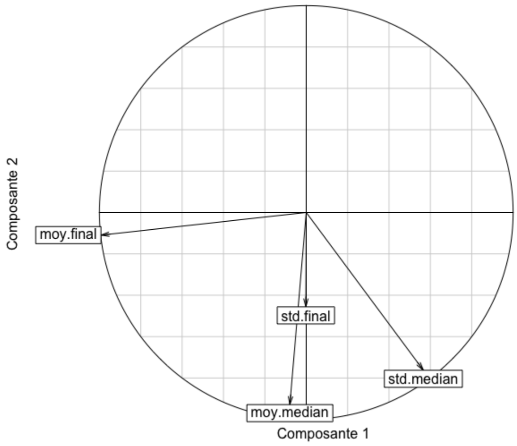


FIGURE 2.2 – Variables dans le premier plan factoriel de l'ACP

2.1.4 Reconstitution de la matrice originelle $\boldsymbol{X}$ centrée en colonne

On obtient C avec la formule $C = XMU$ (X matrice des observations centrées en colonne, $M = I_p$ et U matrice des vecteurs propres représentant les axes factoriels obtenus après ACP).

$$\begin{aligned} & C = XMU \\ \iff \quad & CU^T = XMUU^T \qquad or\ MUU^T = I_p \\ \iff \quad & CU^T = X \end{aligned}$$

La somme $\sum_{\alpha=1}^{k} = c_\alpha u_\alpha^T$ est donc égale à une approximation de la matrice centrée en colonne X d'origine, exprimée en fonction des α premiers axes factoriels de l'ACP. Pour $\alpha = k$, donc en prenant en compte 100% de l'inertie expliquée, on obtient la matrice X (centrée en colonne) d'origine.

2.1.5 Représenter les individus possédant des valeurs manquantes

Nous allons représenter les individus possédant des valeurs manquantes avec la méthode d'imputation par la moyenne : chaque valeur non renseignée sera remplacée par la moyenne de la variable correspondante.

On obtient la matrice (non centrée en colonne) suivante :

	moy.median	std.median	moy.final	std.final
Cor1	10.71	3.90	10.94	4.58
Cor2	12.01	3.71	12.21	4.39
Cor3	10.90	4.01	12.57	3.65
Cor4	10.23	3.04	13.43	4.34
Cor5	10.98	4.41	11.83	3.97
Cor6	11.50	4.30	13.41	4.88
Cor7	10.12	4.03	11.90	4.44
Cor8	10.74	4.65	11.40	4.87

L'ACP effectuée avec la fonction `princomp` donne la représentation suivante des individus sur le premier plan factoriel :

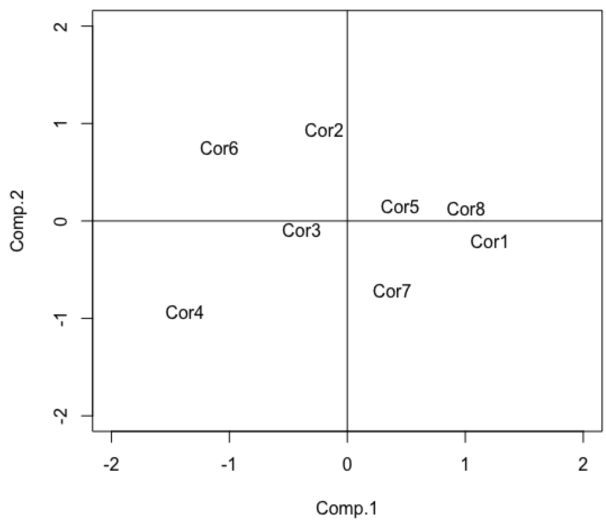


FIGURE 2.3 – Individus (dont les correcteurs 2 et 3 précédemment supprimés) dans le premier plan factoriel de l'ACP

2.2 Utilisation des outils R

L'objectif est de se familiariser avec les fonctions de R permettant d'effectuer une ACP en utilisant le jeu de données de notes du polycopié de cours. Ce jeu de données contient les notes de neuf élèves dans les matières mathématique, sciences, français, latin et dessin.

2.2.1 ACP avec la fonction `princomp`

En effectuant l'ACP avec la fonction `princomp`, on obtient un objet de type `princomp` qui possède plusieurs attributs nous permettant de retrouver les différentes parties d'une ACP.

» **valeurs propres** : correspond à l'attribut `$sdev`, les écart types de chaque composante.

» **vecteurs propres** : correspond à l'attribut `$loadings`.

» **composantes principales** : correspond à l'attribut `$scores`.

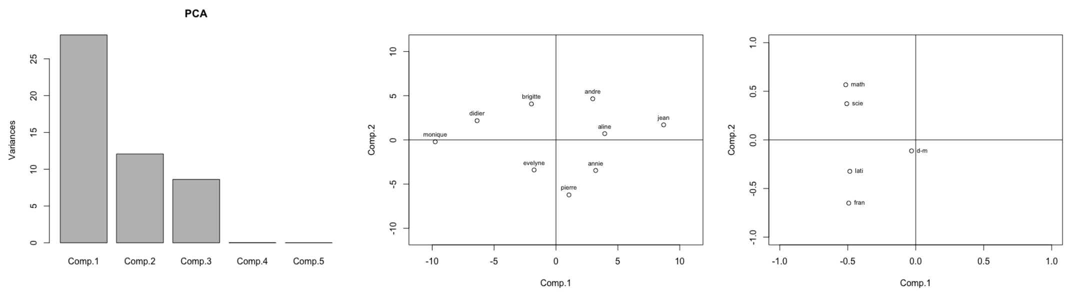
Appeler la méthode `summary` sur le résultat de l'ACP (voir Figure 2.4) donne l'inertie expliquée de chacune des composantes, ainsi que leur pourcentage simple et cumulé.

```
Importance of components:
                        Comp.1 Comp.2 Comp.3    Comp.4    Comp.5
Standard deviation       5.315  3.475  2.935 0.147418 0.099347
Proportion of Variance   0.577  0.247  0.176 0.000444 0.000202
Cumulative Proportion    0.577  0.823  0.999 0.999798 1.000000
```

FIGURE 2.4 – Résultat de `summary` sur l'objet d'ACP retourné par la fonction `princomp`

2.2.2 ACP : fonctions `plot` et `biplot`

La fonction `plot` permet de représenter les résultats de l'ACP de différentes façon. Appeler `plot` sur l'objet ACP donne les pourcentages de variance pour chaque composante (voir Figure 2.5a). Si on appelle la méthode sur `ACP$scores` alors on obtiendra la représentation des individus dans le premier plan factoriel (voir Figure 2.5b). Si on veux utiliser d'autres composantes, il faudra le spécifier directement en choisissant les colonnes à utiliser dans la matrice des composantes principales. Et enfin, si on appelle `plot` sur `ACP$loadings` on obtiendra la représentation des variables d'origine sur le premier plan factoriel (voir Figure 2.5c).



(a) Pourcentages de variance pour chaque composante. (b) Individus sur le premier plan factoriel. (c) Variables d'origine sur le premier plan factoriel.

FIGURE 2.5 – Résultats de la fonction `plot`.

La fonction `biplot` redéfinie pour l'ACP va mélanger les deux représentations sur le premier plan factoriel (par défaut) : on aura les individus ainsi que l'expression de la corrélation entre les anciennes variables et les composantes principales (voir Figure 2.6).

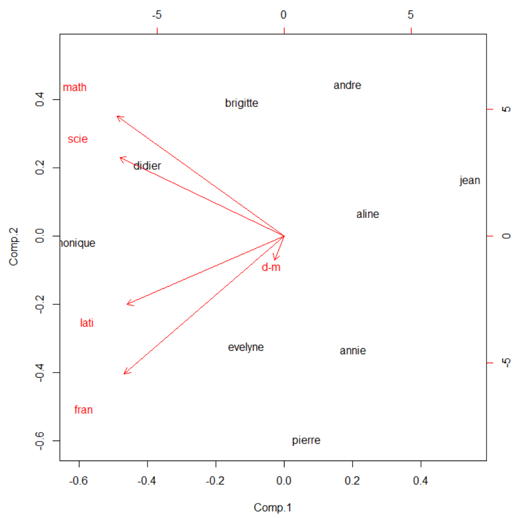


FIGURE 2.6 – Biplot par défaut sur le premier plan factoriel.

On remarque qu'il y a deux échelles de valeur pour chaque axe (haut et bas pour *Comp.1* et gauche et droite pour *Comp.2*). Les axes bas et gauche sont utiliser pour interpréter les corrélations entre les variables d'origine et les nouveaux axes factoriels. Les axes haut et droit mesurent les valeurs prises par les individus dans le nouveau plan factoriel.

Options de la fonction `biplot` :

- **choices** : un vecteur de longueur 2 spécifiant les composantes à utiliser. Par défaut ce sont les deux premières qui sont choisies
- **scale** : réel compris entre 0 et 1, permet d'adapter l'échelle des représentations entre les individus et les variables
  - Les variables sont mise à l'échelle par λ^{scale}
  - Les observations sont mise à l'échelle par $\lambda^{1-scale}$
  - λ représente les valeurs propres calculées par l'ACP
- **pc.biplot** : booléen, si *vrai* utilise la représentation de Gabriel (1971)
  - Les observations sont mises à l'échelle par $\sqrt{n}$ (n est la taille de l'échantillon)
  - Alors le produit interne entre les variables donne une approximation des covariances et les distance entre les observations donnent une approximation de la distance de Mahalanobis

2.3 Données Crabs

2.3.1 ACP sur les données initiales

Voici ci-dessous un tableau récapitulatif de l'ACP effectuée sur les données `crabsquant` sans traitement préalable.

Table 2.1 – ACP sur le jeu de données « crabs » (sans traitement préalable)

	Comp. 1	Comp. 2	Comp. 3	Comp. 4	Comp. 5
Inertie expliquée	11.83	1.14	1	0.37	0.28
Pourcentage d'inertie expliquée	98.25	0.91	0.70	0.09	0.05
Pourcentage d'inertie expliquée cumulé	98.25	99.16	99.86	99.95	100

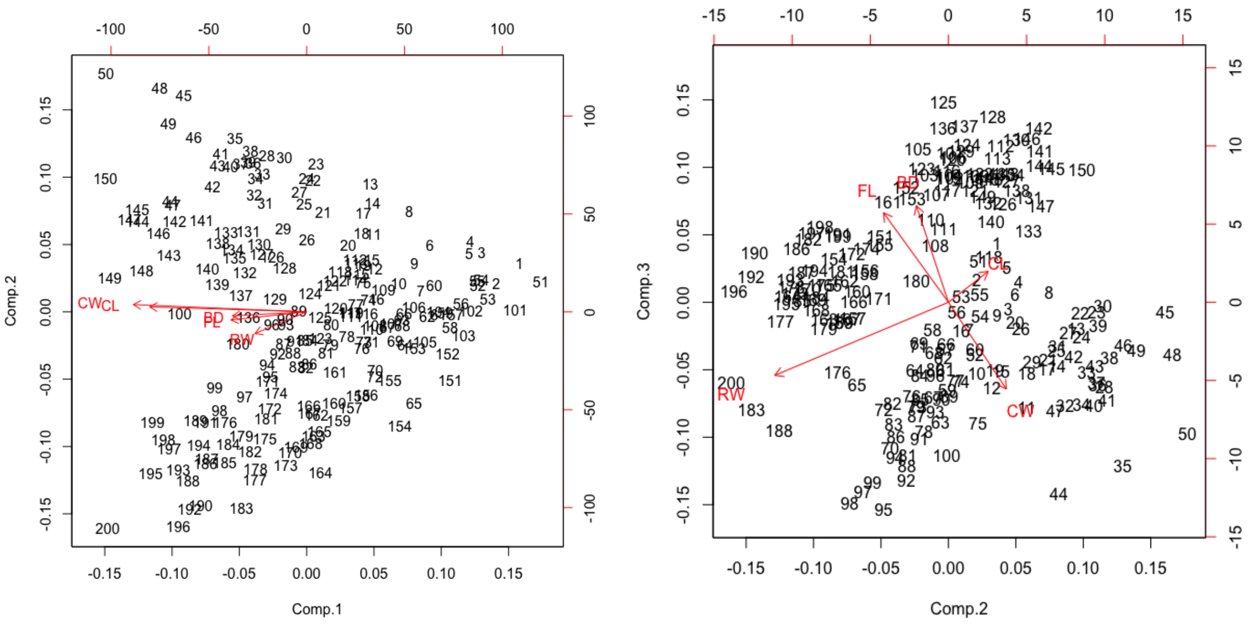
On constate que la première composante représente 98.25% des données à elle toute seule. Si on calcule la matrice de corrélation entre les nouvelles composantes principales et les variables d'origine (Tableau 2.2), on remarque alors que toutes les variables d'origine sont très fortement corrélées à la première composante principale. Tout particulièrement les variables *CL* et *CW*, c'est à dire la longueur et la largeur de la coquille des crabes.

Table 2.2 – Corrélations entre les variables d'origines et les composantes principales (données « crabs »)

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
FL	-0.981	-0.105	0.145	0.077	0.010
RW	-0.909	-0.383	-0.161	-0.021	-0.015
CL	-0.999	0.032	0.025	-0.007	-0.029
CW	-0.997	0.042	-0.062	0.006	0.017
BD	-0.983	-0.053	0.160	-0.068	0.036

On retrouve bien ici l'effet de taille observé précédemment dans la sous-section 1.2.2. La représentation graphique de l'ACP obtenue avec la commande `biplot` (Figure 2.7a) confirme cette corrélation : les variables *CW* et *CL* sont particulièrement bien exprimée par le premier axe factoriel, tout comme

BD et *FL*. Seule la mesure *RW* « s'émancipe » un peu est s'exprime légèrement sur le second axe factoriel.
Une rapide analyse du tableau de corrélation (Tableau 2.2) nous indique d'ailleurs que les composantes 2 et 3 pourraient bien mieux représenter les variables *FL*, *RW* et *BD*. Et effectivement, une représentation des individus sur le plan factoriel créé à partir du deuxième et troisième axe (Figure 2.7b), permet à priori de mieux discriminer la population de crabes selon leur espèce et leur sexe. Mais il ne faut pas oublier qu'en ne considérant pas la premier axe factoriel, on perd 98% de l'inertie expliquée...



(a) ACP crabs : biplot du premier plan factoriel

(b) ACP crabs : biplot du plan factoriel créé à partir des composantes 2 et 3

FIGURE 2.7 – Représentations de l'ACP des données `crabs` sur deux plans factoriels différents

2.3.2 Correction l'effet de taille

Afin de prendre en compte une plus grande partie de l'inertie expliquée, il va nous falloir corriger l'effet de taille. Le principe de l'ACP est d'identifier les axes présentant la plus grande dispersion possible, et l'ACP précédente a bien montré que la taille du crabe est le facteur le plus « dispersant ». Il va donc falloir réduire au maximum cette dispersion afin donner plus d'importance aux autres mesures et peut-être identifier des facteurs morphologiques (autres que la taille) communs au sexe et/ou aux espèces.
On a identifié les variables *CL* et *CW* comme étant les initiatrices principales de cet effet. Il suffit alors de réduire à 0 la dispersion de l'une de ces deux mesures. Une manière simple de faire cela est d'exprimer l'ensemble des autres variables en pourcentage de la mesure dont on veut réduire la dispersion. Nous avons ici choisit de réduire la dispersion de la mesure *CL* (les effets de *CL* et *CW* sur l'ACP sont quasiment identiques, donc ici choisir l'une ou l'autre revient finalement au même, comme on peut le voir dans la Figure 2.8b).
La méthode est donc, pour chaque individu, de diviser toutes les observations par la valeur de *CL* puis de multiplier par 100. Le Tableau 2.3 montre bien le changement se produisant dans les observations.

TABLE 2.3 – Comparatif des observations du premier individu avant et après correction de l'effet de taille

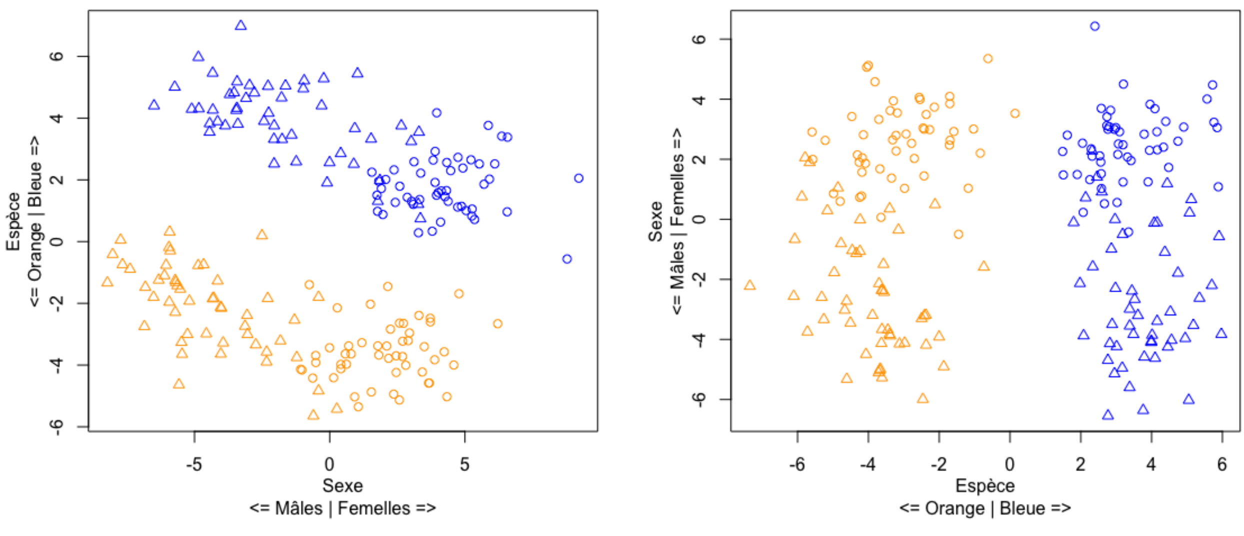
	FL	RW	CL	CW	BD
Avant correction	8.10	6.70	16.10	19.00	7.00
Après correction	50.31	41.61	100.00	118.01	43.48

Tous les individus auront donc maintenant une valeur de 100 pour la variable *CL* : sa dispersion sera nulle. L'ACP effectuée sur les nouvelles données ainsi corrigée (Tableau 2.4) est bien plus « équilibrée » dans le sens où la première composante principale ne représente pas 98% des données mais 53%, et qu'il faut prendre en compte les 3 premiers axes factoriels pour obtenir environ 95% d'inertie expliquée.

TABLE 2.4 – ACP sur le jeu de données `crabs` (avec correction de l'effet de taille)

	Comp. 1	Comp. 2	Comp. 3	Comp. 4	Comp. 5
Inertie expliquée	3.936	3.217	1.333	1.2302	0
Pourcentage d'inertie expliquée	53.2	35.5	6.1	5.19	0
Pourcentage d'inertie expliquée cumulé	53.2	88.7	94.8	100	100

Voici la représentation des données sur le nouveau plan factoriel déterminé par l'ACP. On remarque une séparation très claire du sexe et de l'espèce avec la formation de quatre grappes de points bien séparées.



(a) ACP crabs : correction de l'effet de taille en normalisant par rapport à la variable *CL*

(b) ACP crabs : correction de l'effet de taille en normalisant par rapport à la variable *CW*

FIGURE 2.8 – Représentations de l'ACP des données `crabs` après correction de l'effet de taille

La coloration (bleue ou orange) et la différenciation du type de point (triangle pour les mâles, rond pour les femelles) nous a permis d'identifier la signification des axes du premier plan factoriel : sexe et espèce. Remarquons que choisir *CL* ou *CW* comme base pour la normalisation des observation va simplement inverser les deux premiers axes factoriels.

Conclusion : la suppression de l'effet de taille permet de véritablement discriminer les observations en fonction de l'espèce et du sexe des individus, et non plus seulement en fonction de leur taille.

2.4 Données Pima

On effectue l'analyse en composantes principales sur les variables quantitatives du jeu de données `Pima` et on obtient les résultats suivants.

```
Importance of components:
                        Comp.1 Comp.2  Comp.3 Comp.4 Comp.5 Comp.6   Comp.7
Standard deviation     31.458 13.436 10.6186 9.2122  4.569 2.4770 3.36e-01
Proportion of Variance  0.709  0.129  0.0808 0.0608  0.015 0.0044 8.09e-05
Cumulative Proportion   0.709  0.839  0.9197 0.9806  0.996 0.9999 1.00e+00
```

(a) ACP `Pima` : Pourcentage d'inertie expliquée

```
Loadings:
      Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7
npreg                      -0.165         0.977
glu   -0.981  0.186
bp    -0.109 -0.756  0.239  0.590  0.104
skin         -0.388 -0.754 -0.286  0.437
bmi          -0.228 -0.394        -0.888
ped                                              1.000
age   -0.113 -0.428  0.458 -0.735 -0.100 -0.210
```

(b) ACP `Pima` : Axes factoriels (vecteurs propres résultants de l'ACP)

FIGURE 2.9 – ACP sur `Pima` : inertie expliquée et axes factoriels.

TABLE 2.5 – ACP sur `Pima` : corrélation entre les variables d'origine et les nouveaux axes factoriels.

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7
npreg	-0.16	-0.36	0.32	-0.46	0.01	0.73	0.00
glu	-1.00	0.08	0.01	0.01	0.00	0.00	-0.00
bp	-0.28	-0.83	0.21	0.44	0.04	0.00	0.00
skin	-0.27	-0.50	-0.76	-0.25	0.19	-0.00	-0.00
bmi	-0.29	-0.45	-0.61	0.02	-0.59	0.01	-0.00
ped	-0.17	-0.01	-0.08	-0.07	-0.08	-0.04	0.98
age	-0.33	-0.53	0.45	-0.63	-0.04	-0.05	-0.00

On voit que le premier plan factoriel représente environ 83% de l'inertie expliquée, ce qui est encourageant. Par contre, on remarque aussi que les axes factoriels semblent être particulièrement représentatifs de certaines variables d'origine. Et c'est confirmé par la matrice de corrélation entre les variables d'origine et les nouveaux axes factoriels (voir Tableau 2.5). Le premier axe factoriel est quasiment équivalent à la variable `glu`, le second à la variable *bp*, le troisième représente plutôt les variables `skin` et `bmi`, le quatrième l'âge, le cinquième `bmi` encore et le sixième le nombre de grossesses. Le septième axe est le plus flagrant d'entre tous : il est quasiment équivalent à la variable `ped` d'origine.

L'ACP n'est pas concluante. Les sept composantes principales trouvées sont un reflet trop proche des sept variables initiales. Une représentation graphique des observations sur le premier plan factoriel ne permet pas de distinguer visuellement les groupes de patientes diabétiques et non diabétiques.

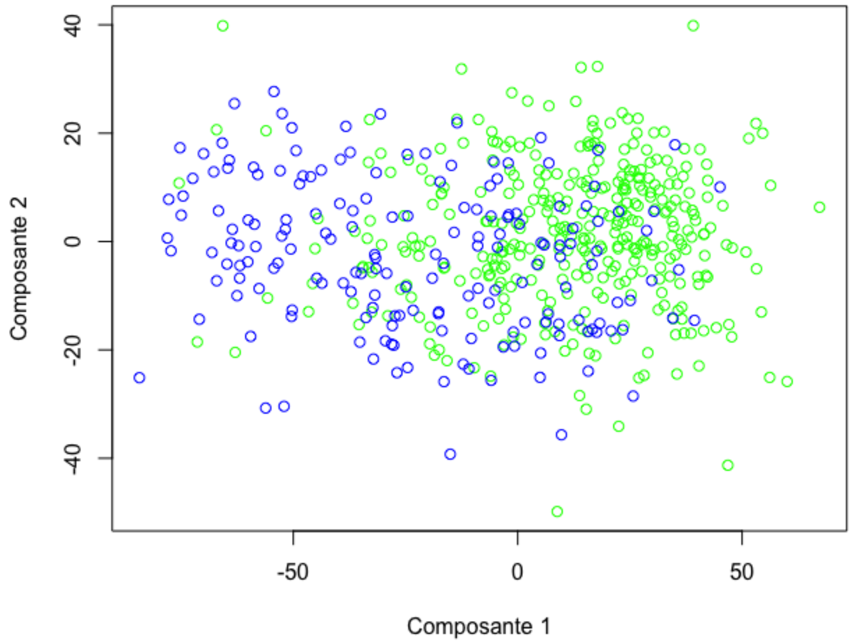


FIGURE 2.10 – Représentation des données `Pima` sur le premier plan factoriel.
Rappel : bleu pour diabétique, vert pour non diabétique

Nous avons fait une nouvelle ACP en changeant la matrice D_p de poids des individus, en spécifiant que le poids d'un individu d'un groupe g valait $\frac{1}{card(g)}$, afin de rééquilibrer le poids de chaque groupe.

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7
Standard deviation	2.202	1.808	1.466	1.324	1.231	0.8801	0.8143
Proportion of Variance	0.324	0.218	0.144	0.117	0.101	0.0517	0.0443
Cumulative Proportion	0.324	0.542	0.686	0.803	0.904	0.9557	1.0000

FIGURE 2.11 – Inertie expliquée de l'ACP avec poids des individus normalisé en fonction de leur groupe

On obtient la table de corrélation suivante entre les variables d'origine est les nouvelles composantes principales :

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7
npreg	-0.54	0.67	-0.01	-0.28	0.20	0.26	0.16
glu	-0.58	-0.10	-0.23	0.71	0.18	0.08	0.06
bp	-0.60	0.05	0.37	0.11	-0.70	-0.01	0.15
skin	-0.64	-0.51	0.27	-0.29	0.30	-0.29	0.24
bmi	-0.62	-0.61	0.27	-0.20	0.07	0.28	-0.32
ped	-0.27	-0.27	-0.83	-0.26	-0.25	-0.01	0.03
age	-0.68	0.58	-0.04	-0.07	0.03	-0.35	-0.31

On remarque que la première composante principale est maintenant corrélée avec toutes les variables d'origine ou presque, la deuxième aussi et la troisièmes dans une moindre mesure.

Au final, ce traitement des données ne change pas grand chose, les classes ne permettent pas de bien différencier les groupes de femmes diabétiques et non diabétiques.

Peut-être serait-il plus pertinent de changer le poids des variables, et d'attribuer plus de poids à celles dont on a vu dans l'analyse descriptive (voir section 1.3) qu'elles ont un lien fort avec le facteur diabète.