

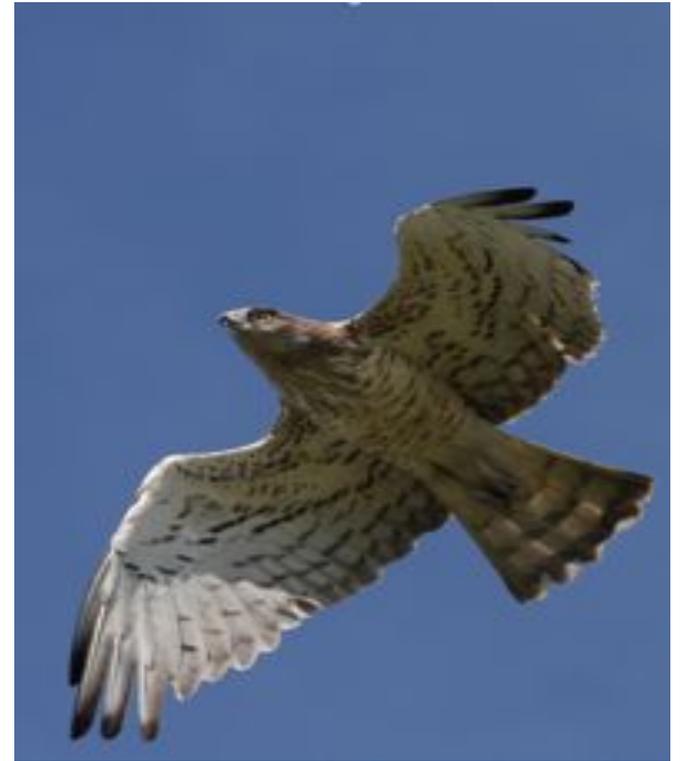
# Détermination du sexe des rapaces à l'aide de l'analyse discriminante

**Vincent Jalby, XLIM**  
**Thérèse Nore, Dpt de Mathématiques**

En collaboration avec

*Jean-Pierre Malafosse, PN des Cévennes*  
*Michael Wink, Université de Heidelberg*

Etude sur l'écologie et la biologie des rapaces,  
notamment la Buse (*Buteo buteo*)  
et le Circaète (*Circaetus gallicus*)



## Que mangent-ils ?



- Quelle est leur longévité?
- Après l'envol, où vont ils?
- A quel âge se reproduisent-ils ?
- Quelle est la taille de leur territoire ?
- Font ils preuve d'agressivité entre eux, ou à l'égard des intrus ?

Cela dépend-il du sexe des oiseaux ?  
Comment déterminer le sexe des oiseaux ?

On remarque parfois un dimorphisme ou un polymorphisme, c'est à dire la présence de plusieurs apparences différentes (phénotypes) chez les individus d'une même espèce.

Ce dimorphisme peut être qualitatif (par exemple lié à la coloration du plumage).

**Comment l'interpréter ?**

Il peut ou non être lié au sexe des oiseaux.



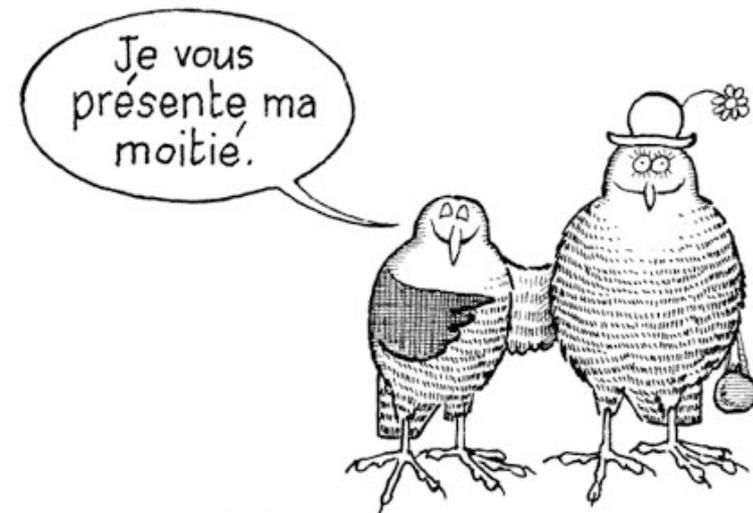
Chez la Buse et le Circaète, pas de dimorphisme qualitatif lié au sexe.



## Il peut aussi y avoir dimorphisme portant sur les mensurations

Chez les mammifères, les mâles sont statistiquement plus grands que les femelles.

Chez les rapaces, c'est le contraire : les femelles sont de plus grand gabarit que les mâles. On parle de « **dimorphisme sexuel inverse** ».



La Hulotte :  
[www.lahulotte.fr](http://www.lahulotte.fr)

Pour certaines espèces, le dimorphisme est si important que la taille des individus permet de différencier les mâles des femelles.



Pour d'autres, chacune des mensurations des femelles est statistiquement supérieure à celle des mâles, mais les deux distributions de valeurs présentent un important recouvrement, de sorte que l'œil humain ne peut déterminer avec certitude le sexe de la plupart des oiseaux.



Cependant, en combinant plusieurs mesures, on peut espérer connaître le sexe d'une bonne proportion d'oiseaux.

C'est le but de l'**Analyse Discriminante**.

Les poussins des deux sexes sont quasiment identiques à la naissance, et la différence de taille s'affirme seulement au cours de la croissance.

Peut-on, à partir des mensurations, et à partir d'un certain âge, prédire le sexe des oiseaux de manière fiable ?

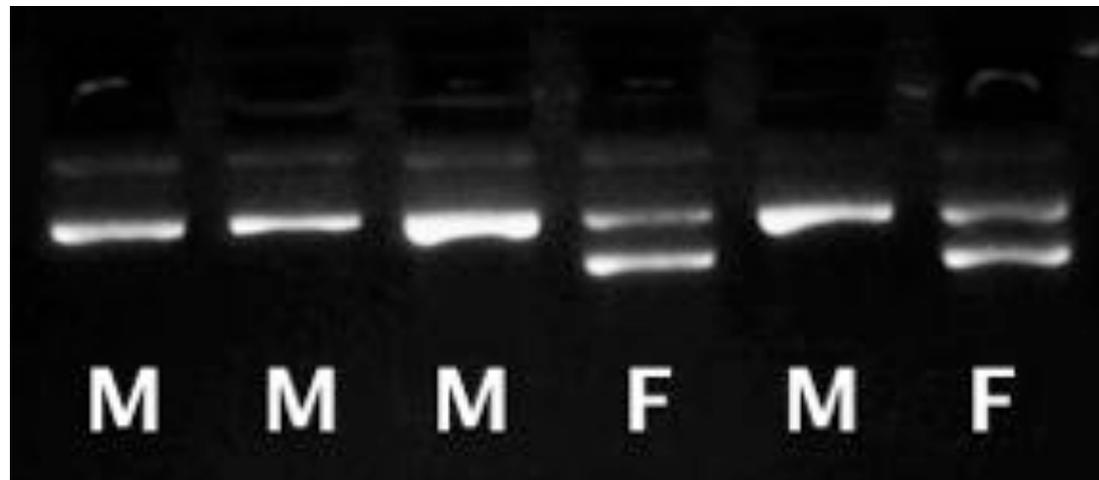


# Analyse ADN

La démarche de l'Analyse Discriminante se base sur deux échantillons d'individus de sexe connu, déterminé par analyse moléculaire.

Chez les oiseaux, les chromosomes sexuels sont ZZ pour les mâles, et WZ pour les femelles.

Une seule barre caractérise les mâles sur le gel d'électrophorèse, tandis qu'on note deux barres chez les femelles.



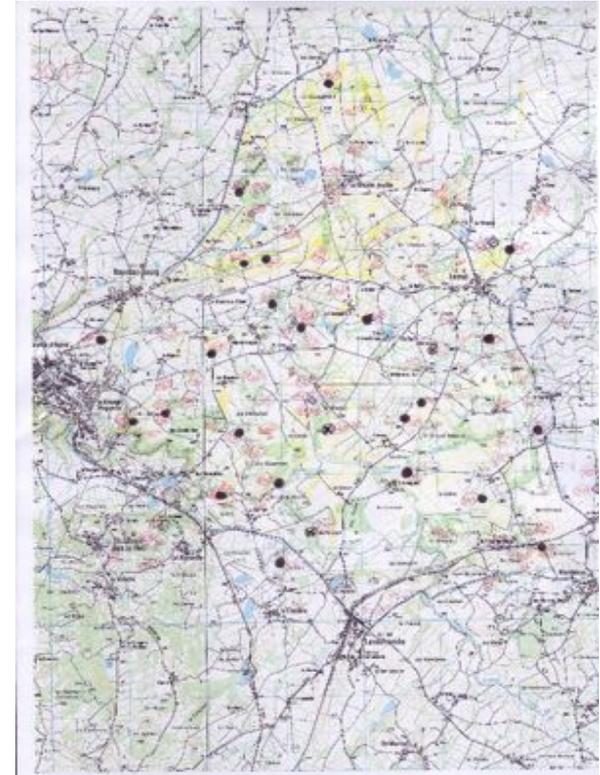
# ECHANTILLONS

Les échantillons utilisés sont constitués si possible de manière aléatoire. Nous procédons par recherche et suivi systématique de tous les nids sur les zones d'étude.

C'est la zone qui est « échantillon aléatoire » !



Prospection systématique par observation  
Dans le Parc National des Cévennes.



Prospection systématique d'une  
zone échantillon pour trouver  
Tous les nids de Buse.

Les nids, situés haut dans les arbres, sont parfois difficiles d'accès, et les manipulations, l'enregistrement des données, sont délicats à réaliser....

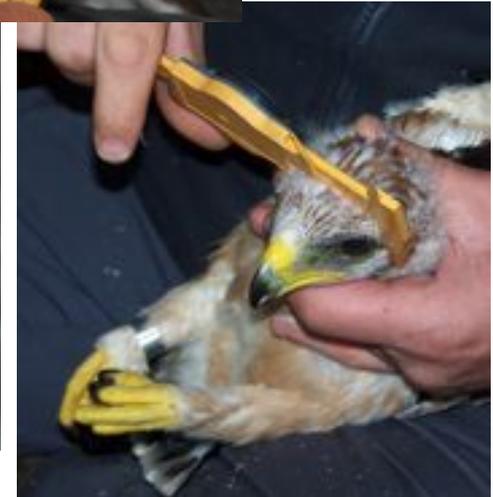


# MESURES

Sur chaque poussin manipulé au nid, nous relevons :



La longueur de l'aile (**AP**), et la pousse des 4<sup>e</sup> et 5<sup>e</sup> rémige primaire (**RP4** et **RP5**)



La longueur de la tête (**LT**), la largeur de la tête (**LLT**), et certaines dimensions du bec (**LBC**).

## Le poids (**P**)



Les dimensions latérales du tarse : petit et grand diamètres (**PD**, **GD**) la longueur du tarse (**T**).

# ANALYSE DISCRIMINANTE

Construire une **fonction affine** des mensurations des oiseaux :

$$D = \sum a_i X_i + \text{cste}$$

permettant de **discriminer** les mâles des femelles.

La fonction D doit donc

- prendre des valeurs différentes entre les groupes
- prendre des valeurs proches dans chaque groupe

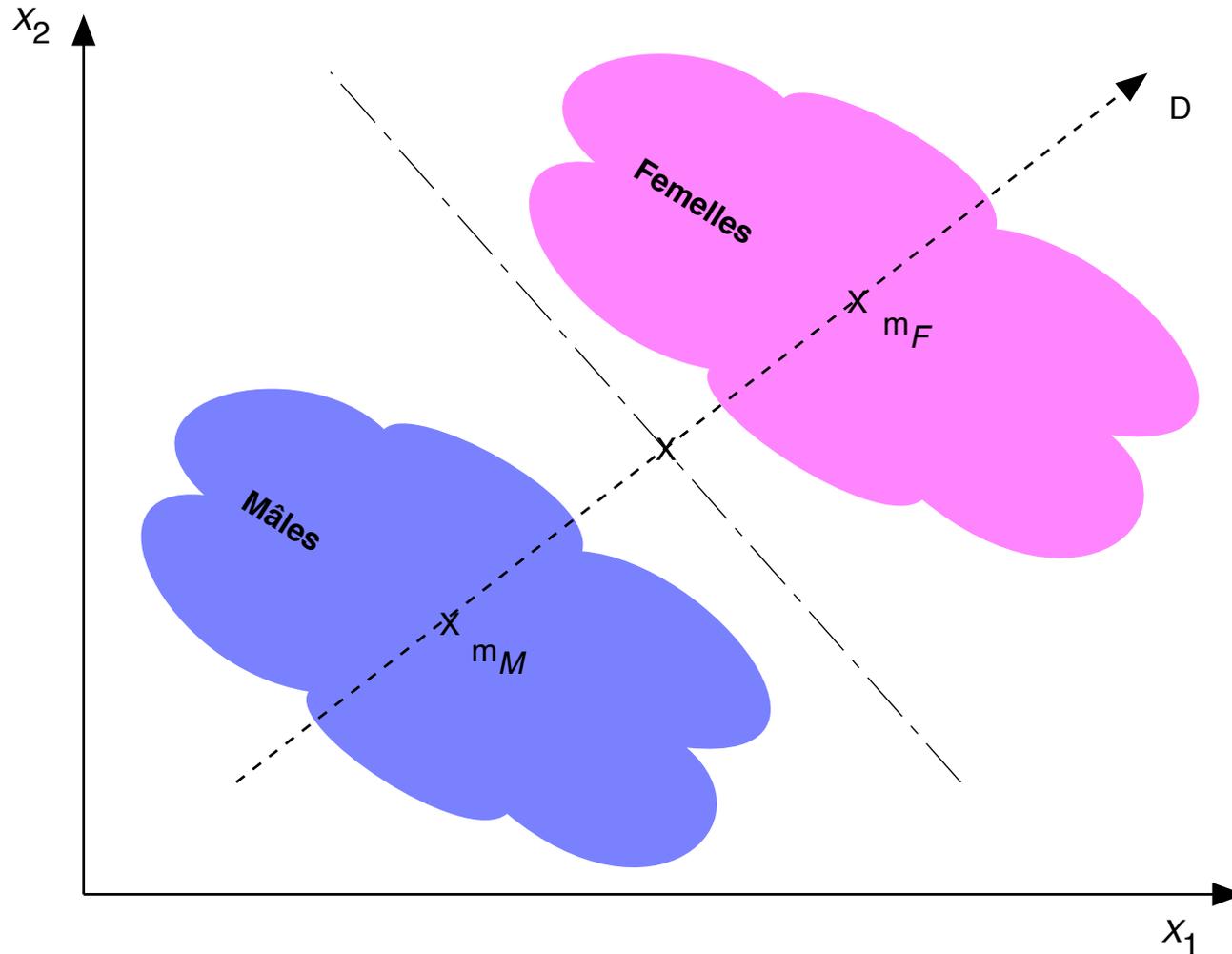
# ANALYSE DISCRIMINANTE

La fonction D doit maximiser le rapport

Variance Intergroupe (D) / Variance Intragroupe (D)

Mathématiquement, cela se ramène à un problème d'optimisation sous contrainte que l'on résout à l'aide de la méthode de Lagrange !

# ANALYSE DISCRIMINANTE



# ANALYSE DISCRIMINANTE

Chaque oiseau est alors classé dans le groupe dont le centre est le plus proche :

$$|d-m_M| < |d-m_F] \rightarrow \text{M\^a}le$$

Le % d'oiseaux bien classés permet de mesurer la qualité de la fonction discriminante.

# ANALYSE DISCRIMINANTE

## Difficultés :

- Limiter le nombre de variables intervenant dans la fonction discriminante  $D$
- Données manquantes sur certaines variables
- Corrélation entre les variables

# TAILLE DES ECHANTILLONS

Poussins d'Autour des palombes		
	Tous	AP $\geq$ 180
Male	239	60
Femelle	187	53
<b>Total</b>	<b>426</b>	<b>113</b>

# TAILLE DES ECHANTILLONS

Poussins de Buse variable		
	Tous	AP $\geq$ 205
Male	71	43
Femelle	108	66
<b>Total</b>	<b>179</b>	<b>109</b>

# TAILLE DES ECHANTILLONS

Poussins de Circaète Jean-le-Blanc		
	Tous	AP $\geq$ 170
Male	59	45
Femelle	47	33
<b>Total</b>	<b>106</b>	<b>78</b>

# DONNÉES MANQUANTES

<b>Poussins d'Autour des palombes</b>	
<b>Echantillon</b>	<b>113</b>
Masse (P)	110
Tarse (T)	105
Petit diamètre (PD)	25
Grand diamètre (GD)	23
<b><i>Toutes les variables</i></b>	<b>23</b>

# DONNÉES MANQUANTES

<b>Poussins de Buse variable</b>	
<b>Echantillon</b>	<b>109</b>
RP4 / RP5	108
Masse (P)	107
Tarse (T)	107
PD / GD	108
Hauteur Bec (HB)	80
<b><i>Toutes les variables</i></b>	<b><i>74</i></b>

# DONNÉES MANQUANTES

<b>Poussins de Circaètes</b>	
<b>Echantillon</b>	<b>78</b>
PD / GD	78
Tarse (T)	76
Hauteur Bec (HB)	68
RP4 / RP5	61
<b><i>Toutes les variables</i></b>	<b>53</b>

# TESTS STATISTIQUES

Un test statistique permet de trancher entre deux **hypothèses** portant sur un **paramètre** en fonction d'observations faites sur un échantillon.

Les hypothèses sont du type

$$H_0 : \Theta = 0 \text{ contre } H_1 : \Theta > 0$$

On utilise pour cela une statistique  $T$  dont on connaît la loi, la plupart du temps, un estimateur de  $\Theta$ .

# TESTS STATISTIQUES

On accepte l'hypothèse  $H_1$  lorsque l'observation  $t$  de  $T$  est supérieure à une valeur limite  $t_\alpha$  déterminée par un risque  $\alpha$ , i.e.,  $P(T > t_\alpha) = \alpha$ . La **règle de décision** est :

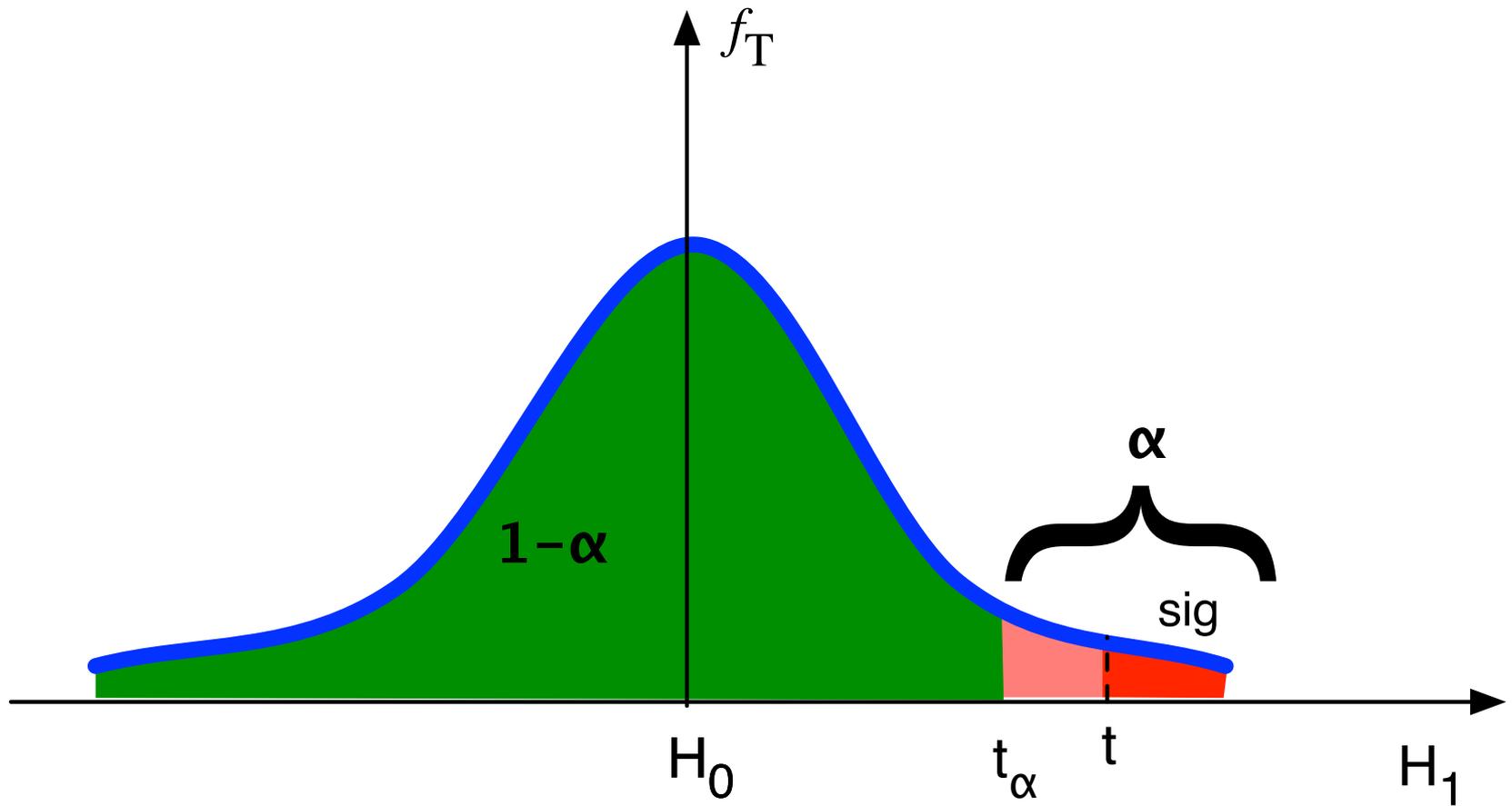
$$t > t_\alpha \rightarrow \text{on accepte } H_1$$

On peut aussi utiliser la **probabilité critique** ou **signification** :

$$\text{sig} = P(T > t) < \alpha \rightarrow \text{on accepte } H_1$$

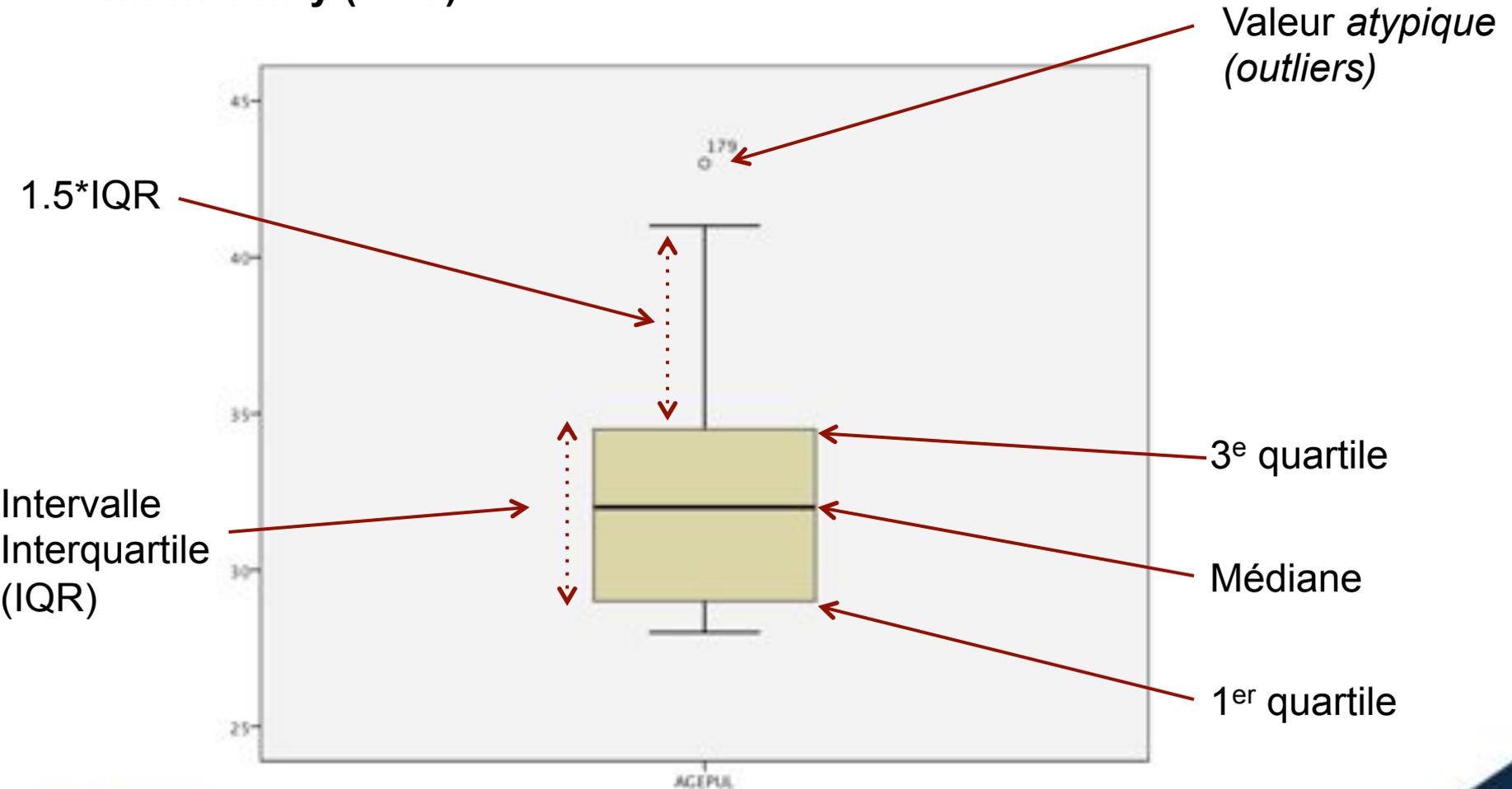
Dans la pratique, on prend  $\alpha = 5 \%$ .

# TESTS STATISTIQUES



# BOITE À MOUSTACHES (BOX-PLOT)

John W. Tukey (1977)

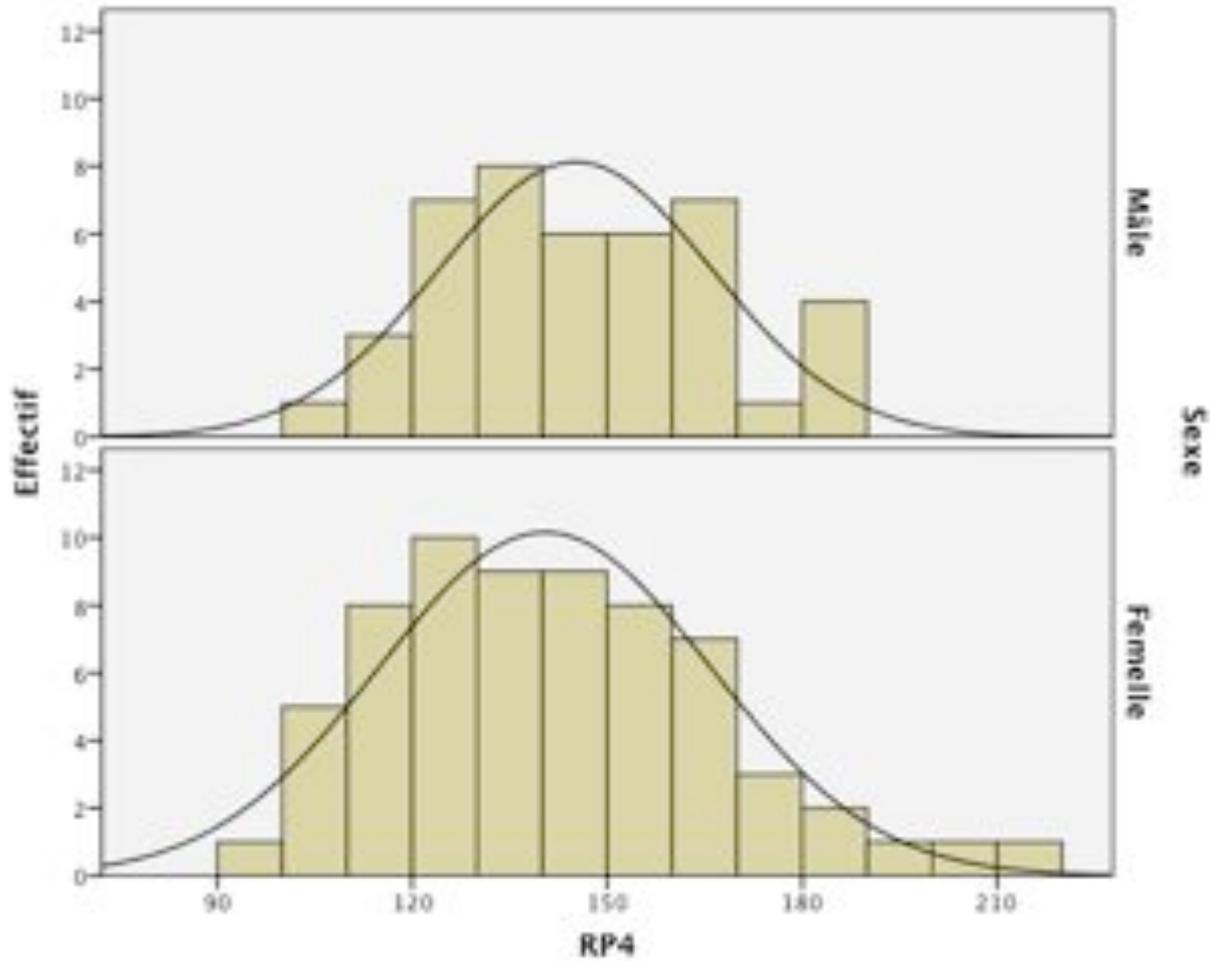


# Normalité

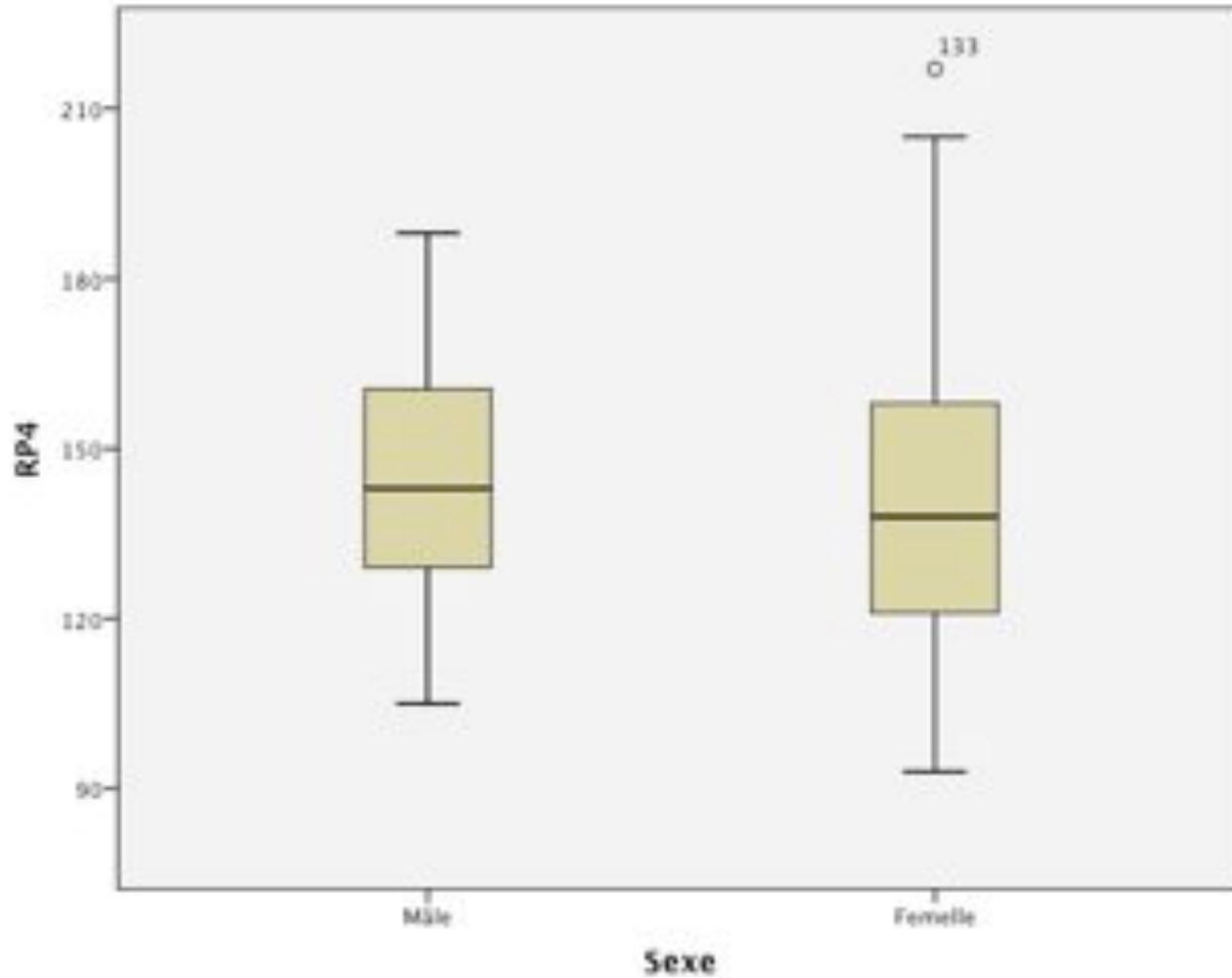
Afin de pouvoir utiliser les techniques statistiques classiques, il est indispensable de vérifier que les variables étudiées suivent (approximativement) des lois normales.

Dans la suite, on se limitera aux poussins de buse.

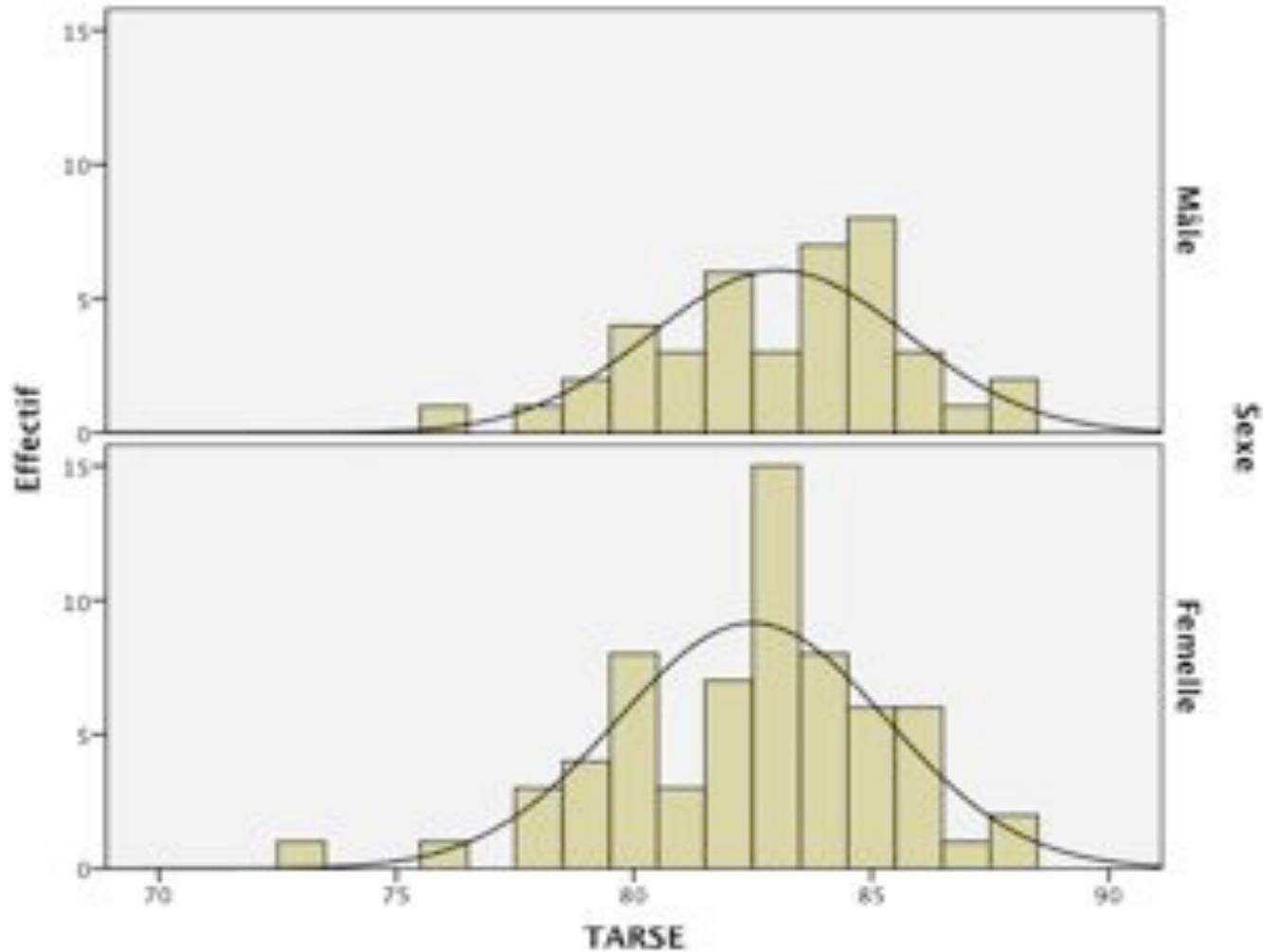
# NORMALITÉ



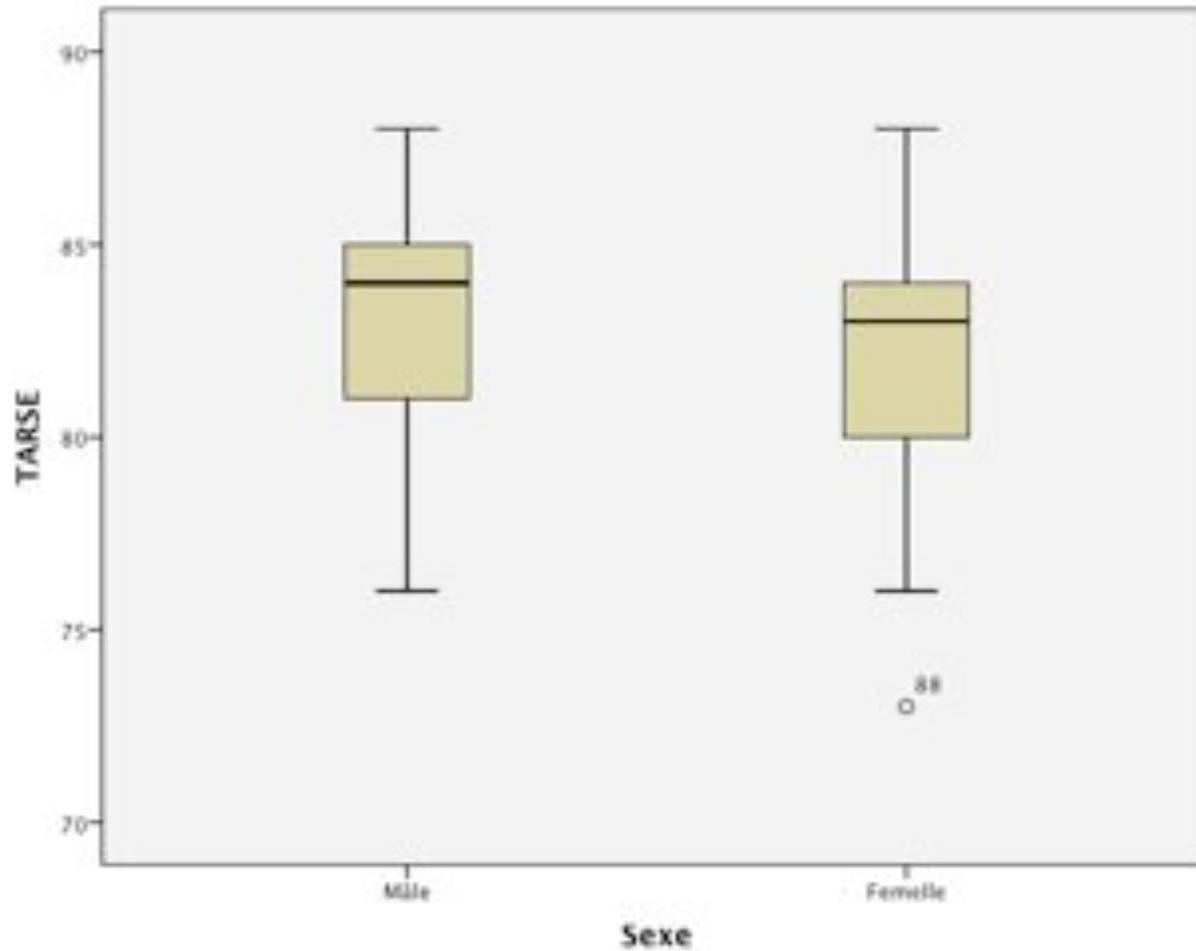
# NORMALITÉ



# NORMALITÉ



# NORMALITÉ



# NORMALITÉ

**asymétrie** (Skewness) :

$$B_1 = \frac{1}{n} \sum_{i=1}^n \left( \frac{x_i - \bar{x}}{s} \right)^3$$

**aplatissement** (Kurtosis) :

$$B_2 = \frac{1}{n} \sum_{i=1}^n \left( \frac{x_i - \bar{x}}{s} \right)^4 - 3$$

# NORMALITÉ

La statistique de **Jarque-Bera**

$$JB = \frac{n}{6} \left( B_1^2 + \frac{B_2^2}{4} \right)$$

suit une loi du  $\chi^2(2)$  permettant d'obtenir un test.

# NORMALITÉ

Variable	Sexe	Effectif	Asymétrie	Applatissement	JB	Signification
RP4	Mâle	41	0,322	-0,631	1,389	0,499
	Femelle	65	0,616	0,403	4,551	0,103
Tarse	Mâle	41	-0,414	-0,044	1,175	0,556
	Femelle	65	-0,635	0,968	6,906	<b>0,032</b>

# NORMALITÉ

On peut aussi utiliser la **fonction de répartition empirique** (fréquences cumulées) de l'échantillon :

$$F_n(x) = \frac{1}{n} \text{Card}(x_i : x_i \leq x)$$

et la comparer à la fonction de répartition de la loi normale.

C'est le test de **Kolmogorov-Smirnov (KS)** et la **droite de Henry (QQ-Plot)**.

# NORMALITÉ

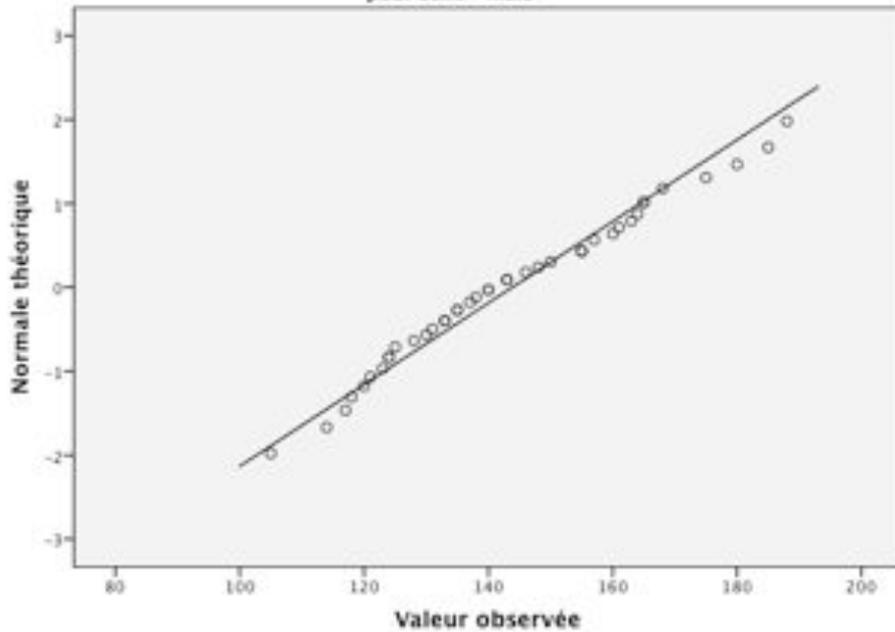
La **droite de Henry** (ou Q-Q plot) consiste à représenter les points

$$\left( x_i, F^{-1}(F_n(x_i)) \right)$$

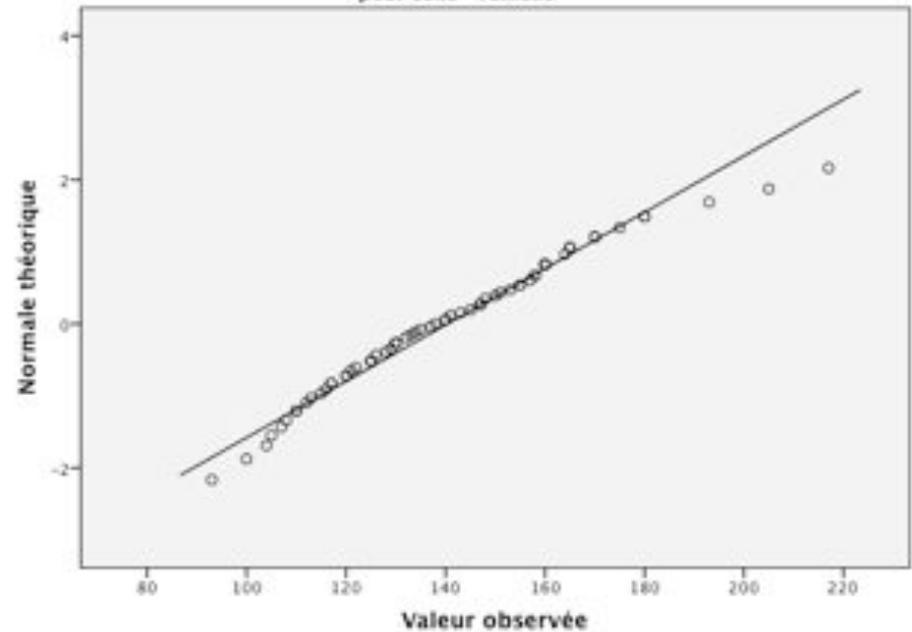
qui sont alignés lorsque l'échantillon est issu d'une loi normale.

# NORMALITÉ

Normogramme Q-Q des résidus de RP4  
pour Sexe = Mâle



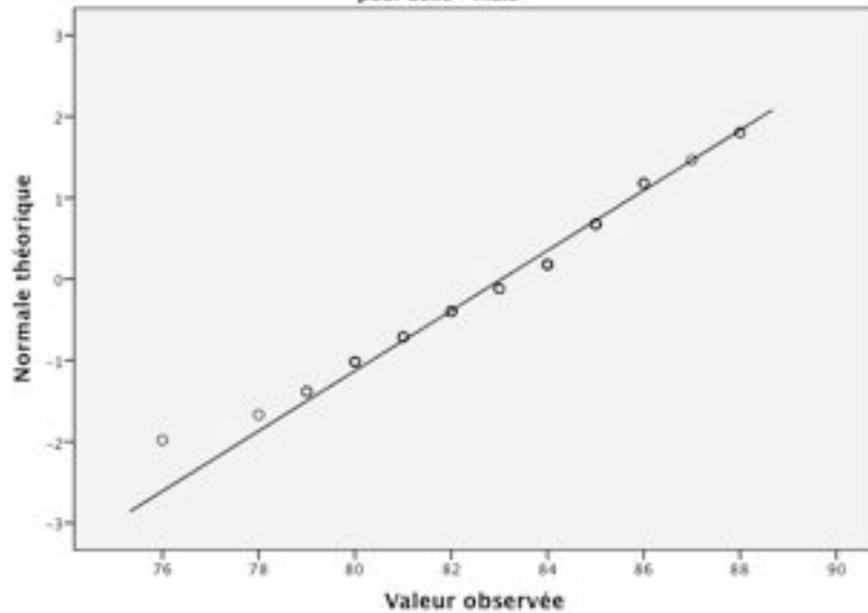
Normogramme Q-Q des résidus de RP4  
pour Sexe = Femelle



# NORMALITÉ

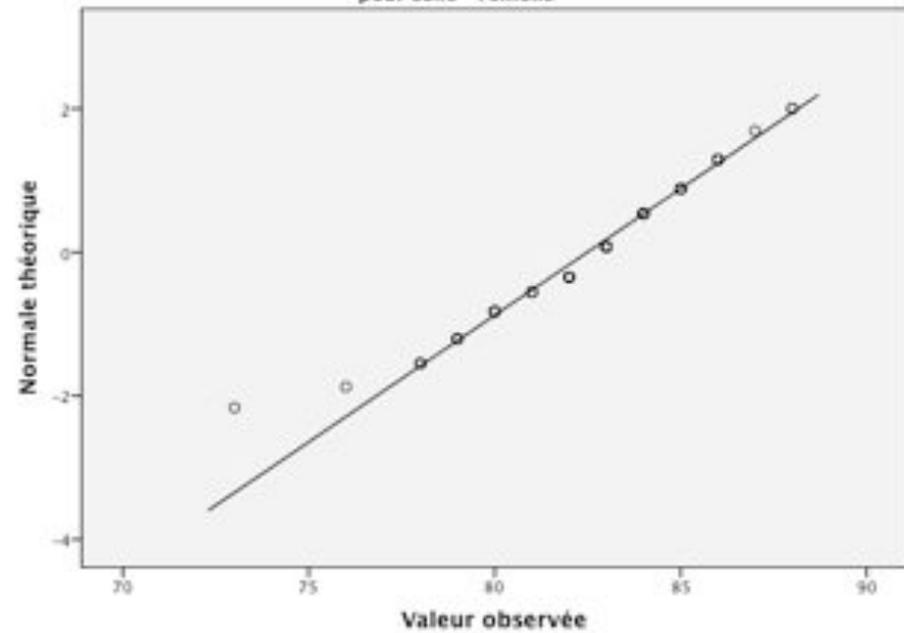
Normogramme Q-Q des résidus de TARSE

pour Sexe= Mâle



Normogramme Q-Q des résidus de TARSE

pour Sexe= Femelle



# NORMALITÉ

Le test de **Kolmogorov-Smirnov (KS)** est basé sur la statistique

$$D_n = \sup_x |F_n(x) - F(x)|$$

qui est nulle lorsque l'échantillon est issu d'une loi normale.

Le test de **Shapiro-Wilk** est une approche non-paramétrique du test de normalité.

# NORMALITÉ

Tests de normalité

Sexe		Kolmogorov-Smirnov <sup>a</sup>			Shapiro-Wilk		
		Statistique	ddl	Signification	Statistique	ddl	Signification
RP4	Mâle	,086	41	,200*	,974	41	,474
	Femelle	,073	65	,200*	,972	65	,146
TARSE	Mâle	,150	41	,021	,969	41	,319
	Femelle	,156	65	,000	,961	65	,037

\*. Il s'agit d'une borne inférieure de la signification réelle.

a. Correction de signification de Lilliefors

# CORRÉLATIONS

Afin d'exclure certaines variables trop corrélées à l'âge des oiseaux et donc non exploitables dans la discrimination, nous avons étudié les corrélations entre toutes les variables.

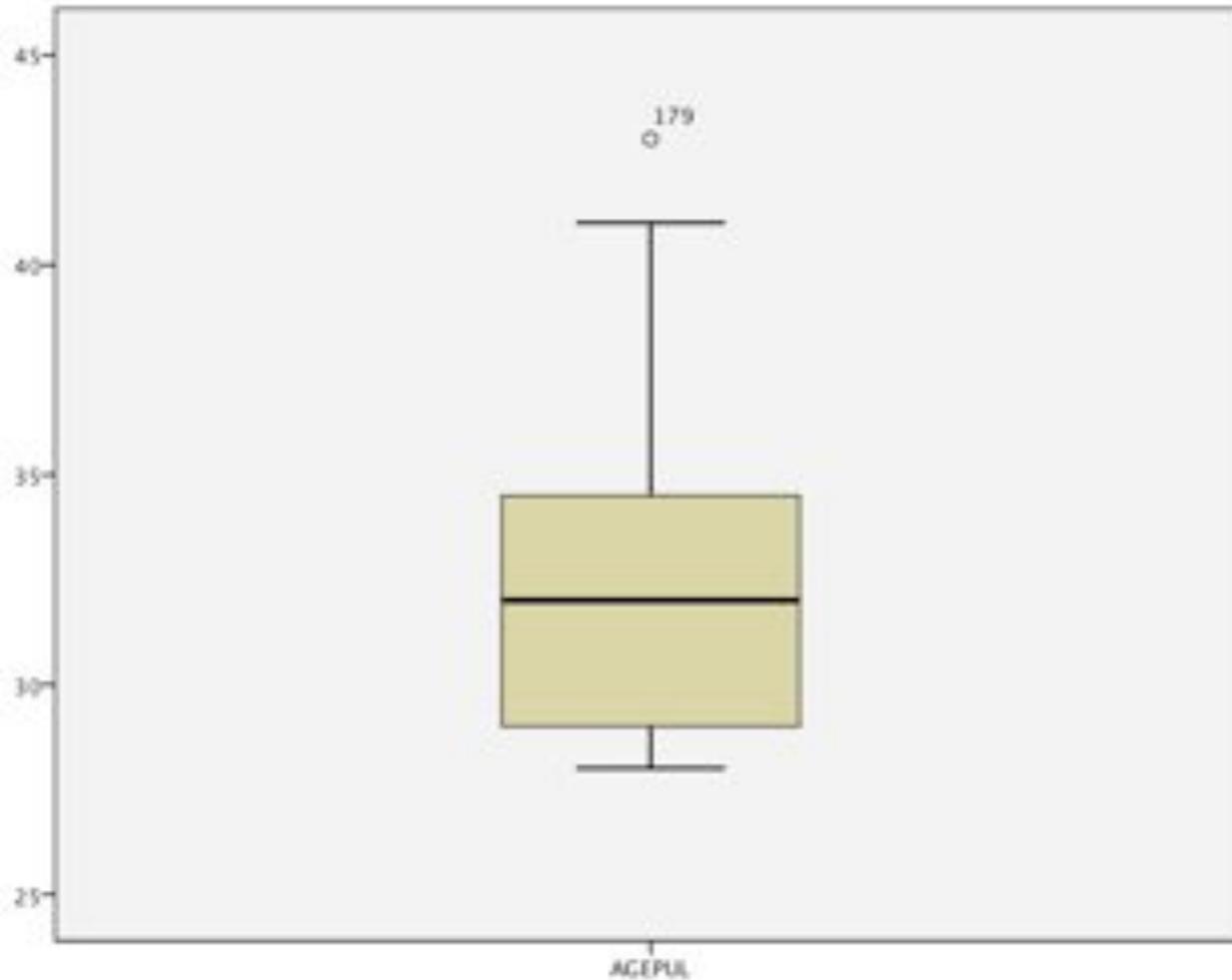
# CORRÉLATIONS

Corrélations	AGE	POIDS	RP4	RP5	TARSE	PD	GD	LBC
AGE	1	0,287	0,895	0,899	0,276	-0,047	0,077	0,356
POIDS	0,287	1	0,361	0,367	0,384	0,349	0,452	0,433
RP4	0,895	0,361	1	0,988	0,432	-0,082	0,063	0,375
RP5	0,899	0,367	0,988	1	0,409	-0,081	0,062	0,37
TARSE	0,276	0,384	0,432	0,409	1	-0,089	0,029	0,19
PD	-0,047	0,349	-0,082	-0,081	-0,089	1	0,675	0,302
GD	0,077	0,452	0,063	0,062	0,029	0,675	1	0,264
LBC	0,356	0,433	0,375	0,37	0,19	0,302	0,264	1

# CORRÉLATIONS

<i>Corrélations intra-groupes combinées</i>								
	AGE	POIDS	RP4	RP5	TARSE	PD	GD	LBC
AGE	1	0,307	0,896	0,902	0,263	-0,05	0,119	0,427
POIDS	0,307	1	0,39	0,389	0,404	0,178	0,362	0,353
RP4	0,896	0,39	1	0,987	0,402	-0,051	0,132	0,46
RP5	0,902	0,389	0,987	1	0,377	-0,06	0,141	0,443
TARSE	0,263	0,404	0,402	0,377	1	-0,035	0,048	0,27
PD	-0,05	0,178	-0,051	-0,06	-0,035	1	0,563	0,039
GD	0,119	0,362	0,132	0,141	0,048	0,563	1	0,095
LBC	0,427	0,353	0,46	0,443	0,27	0,039	0,095	1

# CORRÉLATIONS



# ANALYSE DE LA VARIANCE (ANOVA)

En raison de nombreuses données manquantes, nous souhaitons supprimer les variables qui n'ont qu'un faible pouvoir discriminant.

Pour cela, nous utilisons une analyse de la variance.

# ANALYSE DE LA VARIANCE (ANOVA)

Elle repose sur une décomposition de la variance d'une variable en variance « intra-groupe » (W) et variance « inter-groupe » (B) :

$$\sum_{ij} (x_{ij} - \bar{x})^2 = \sum_i n_i (\bar{x}_i - \bar{x})^2 + \sum_{ij} (x_{ij} - \bar{x}_i)^2$$

$$V = B + W$$

# ANALYSE DE LA VARIANCE (ANOVA)

La variable sera d'autant plus discriminante que le rapport B/W sera important.

Pour effectuer un test, on utilise la statistique suivante :

$$F = \frac{B/(2 - 1)}{W/(n - 2)} \hookrightarrow F(2 - 1, n - 2)$$

# ANALYSE DE LA VARIANCE (ANOVA)

		Somme des carrés	ddl	Moyenne des carrés	F	Signification
POIDS	Inter-groupes	88006,648	1	88006,648	10,841	,001
	Intra-groupes	844269,767	104	8117,979		
	Total	932276,415	105			
TARSE	Inter-groupes	7,785	1	7,785	1,004	,319
	Intra-groupes	806,149	104	7,751		
	Total	813,934	105			
PD	Inter-groupes	9,801	1	9,801	84,425	,000
	Intra-groupes	12,189	105	,116		
	Total	21,990	106			
GD	Inter-groupes	9,437	1	9,437	40,163	,000
	Intra-groupes	24,672	105	,235		
	Total	34,109	106			
LBC	Inter-groupes	20,042	1	20,042	19,833	,000
	Intra-groupes	105,099	104	1,011		
	Total	125,142	105			

# ANALYSE DE LA VARIANCE (ANOVA)

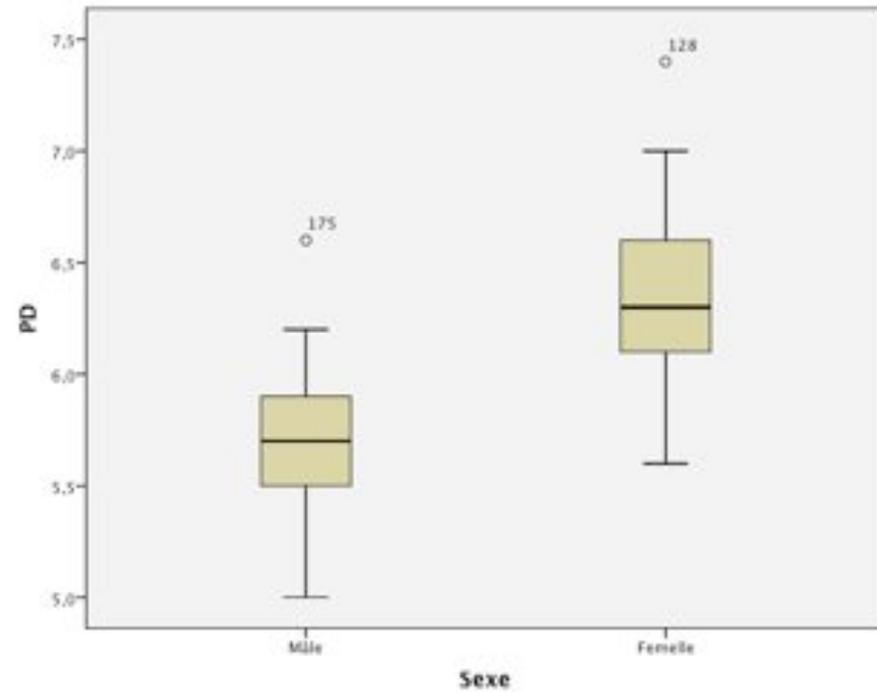
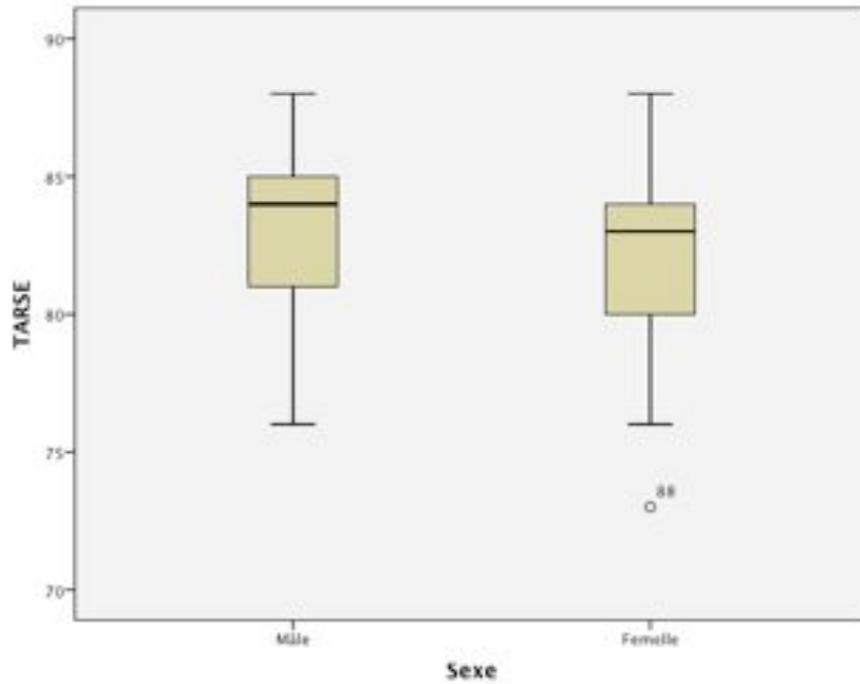
Test d'homogénéité des variances

	Statistique de Levene	ddl1	ddl2	Signification
POIDS	1,498	1	104	,224
TARSE	,000	1	104	,999
PD	1,320	1	105	,253
GD	,121	1	105	,729
LBC	,198	1	104	,658

# ANALYSE DE LA VARIANCE (ANOVA)

		N	Moyenne	Ecart-type	Erreur standard	Intervalle de confiance à 95% pour la moyenne	
						Borne inférieure	Borne supérieure
POIDS	Mâle	43	724,65	84,835	12,937	698,54	750,76
	Femelle	63	783,33	93,498	11,780	759,79	806,88
	Total	106	759,53	94,228	9,152	741,38	777,68
TARSE	Mâle	41	83,05	2,701	,422	82,20	83,90
	Femelle	65	82,49	2,835	,352	81,79	83,19
	Total	106	82,71	2,784	,270	82,17	83,24
PD	Mâle	43	5,714	,3121	,0476	5,618	5,810
	Femelle	64	6,331	,3585	,0448	6,242	6,421
	Total	107	6,083	,4555	,0440	5,996	6,170
GD	Mâle	43	7,535	,4942	,0754	7,383	7,687
	Femelle	64	8,141	,4783	,0598	8,021	8,260
	Total	107	7,897	,5673	,0548	7,788	8,006
LBC	Mâle	43	18,86	1,060	,162	18,53	19,19
	Femelle	63	19,75	,967	,122	19,50	19,99
	Total	106	19,39	1,092	,106	19,18	19,60

# ANALYSE DE LA VARIANCE (ANOVA)



# RÉSULTATS

## Poussins d'autour

$$D = 0,013 P + 0,089 T - 17,570$$

Matrice de structure

	Fonction
	1
Poids	,938
Tarse	,546

Fonctions aux barycentres des groupes

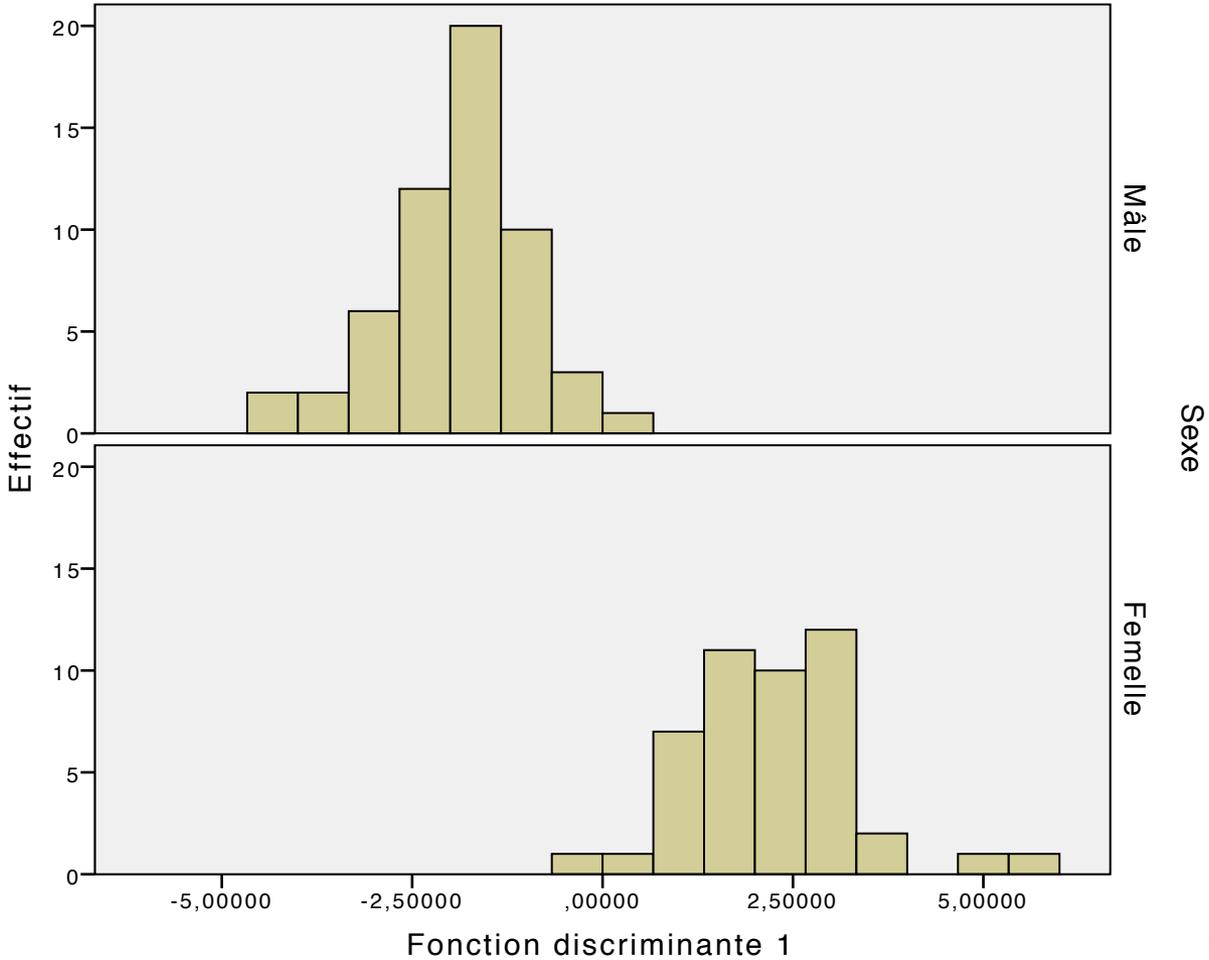
	Fonction
Sexe	1
Mâle	-1,883
Femelle	2,292

Valeurs propres

Fonction	Valeur propre	% de la variance	% cumulé	Corrélation canonique
1	4,402 <sup>a</sup>	100,0	100,0	,903

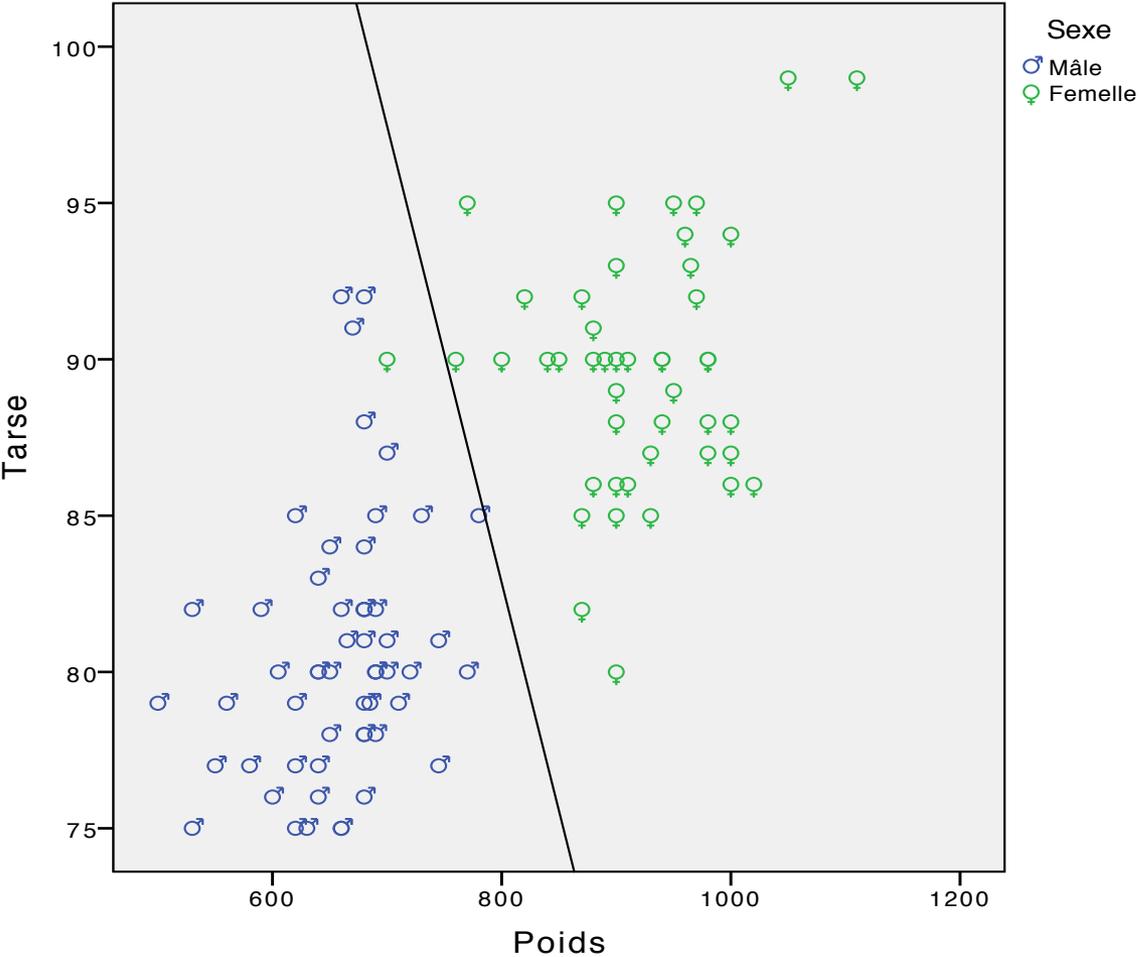
# RÉSULTATS

## Poussins d'autour



# RÉSULTATS

## Poussins d'autour



# RÉSULTATS

## Poussins d'autour

### Résultats du classement<sup>a</sup>

			Classe(s) d'affectation prévue (s)		Total
			Mâle	Femelle	
Original	Effectif	Mâle	56	0	56
		Femelle	1	45	46
	%	Mâle	100,0	,0	100,0
		Femelle	2,2	97,8	100,0

a. 99,0% des observations originales classées correctement.

# RÉSULTATS

## Poussins de buse variable

$$D = 2,645 \text{ PD} + 0,400 \text{ LBC} - 23,847$$

Matrice de structure

	Fonction
	1
PD	,917
BEC	,444

Fonctions aux barycentres des groupes

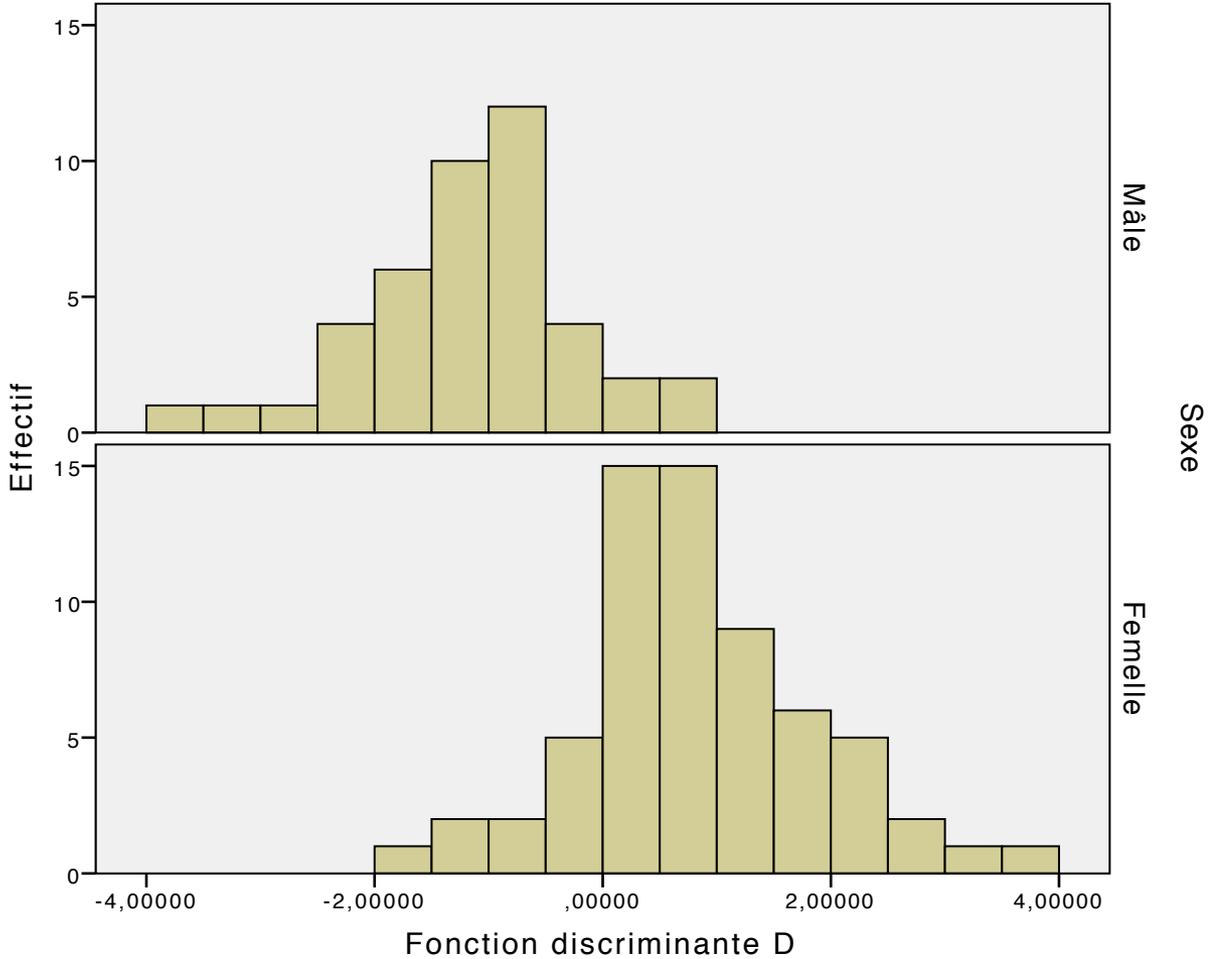
	Fonction
Sexe	1
m	-1,197
f	,804

Valeurs propres

Fonction	Valeur propre	% de la variance	% cumulé	Corrélation canonique
1	,980 <sup>a</sup>	100,0	100,0	,704

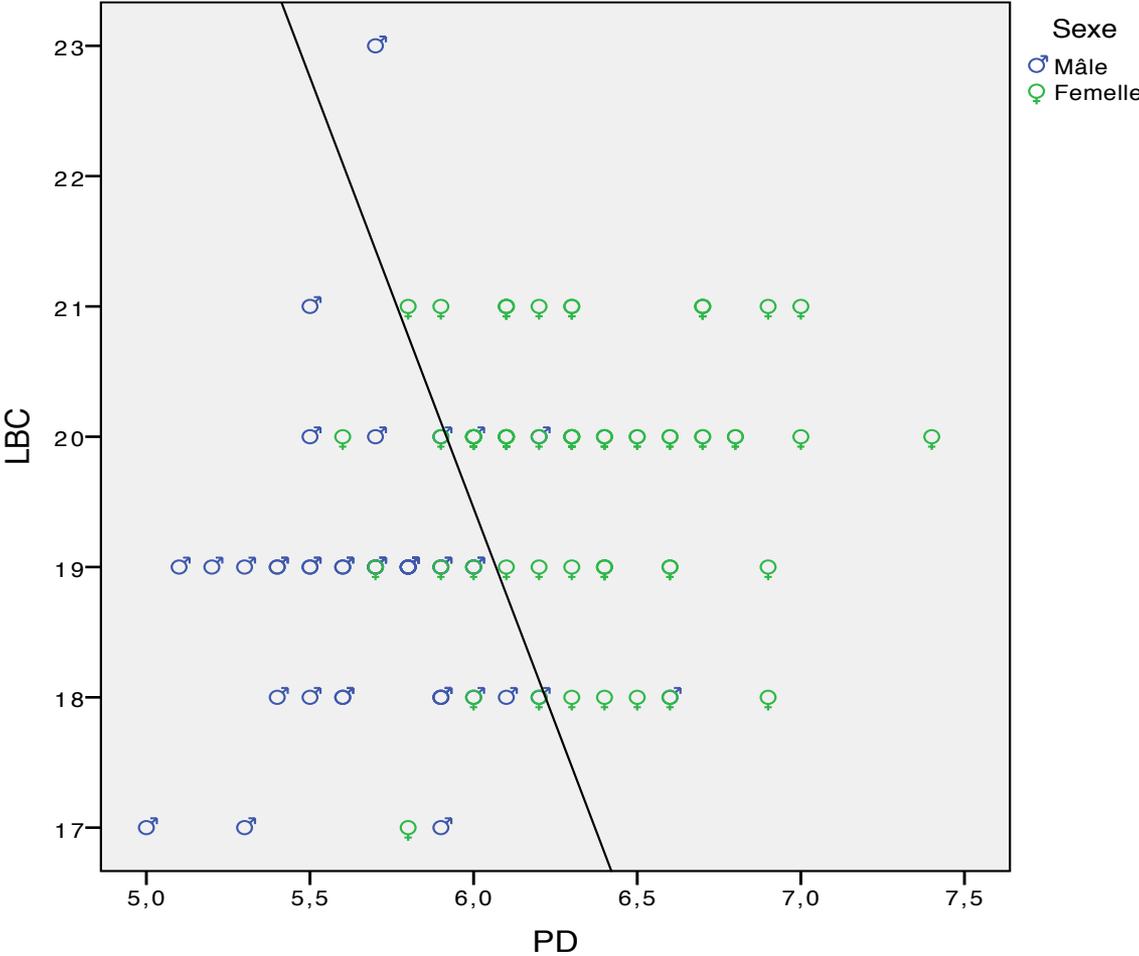
# RÉSULTATS

## Poussins de buse variable



# RÉSULTATS

## Poussins de buse variable



# RÉSULTATS

## Poussins de buse variable

Résultats du classement<sup>a</sup>

			Classe(s) d'affectation prévue (s)		Total
			m	f	
Original	Effectif	m	39	4	43
		f	8	56	64
	%	m	90,7	9,3	100,0
		f	12,5	87,5	100,0

a. 88,8% des observations originales classées correctement.

# RÉSULTATS

## Poussins de Circaète

$$D = 1,013 \text{ GD} + 0,872 \text{ PD} - 20,962$$

Matrice de structure

	Fonction
	1
GD	,898
PD	,760

Fonctions aux barycentres des groupes

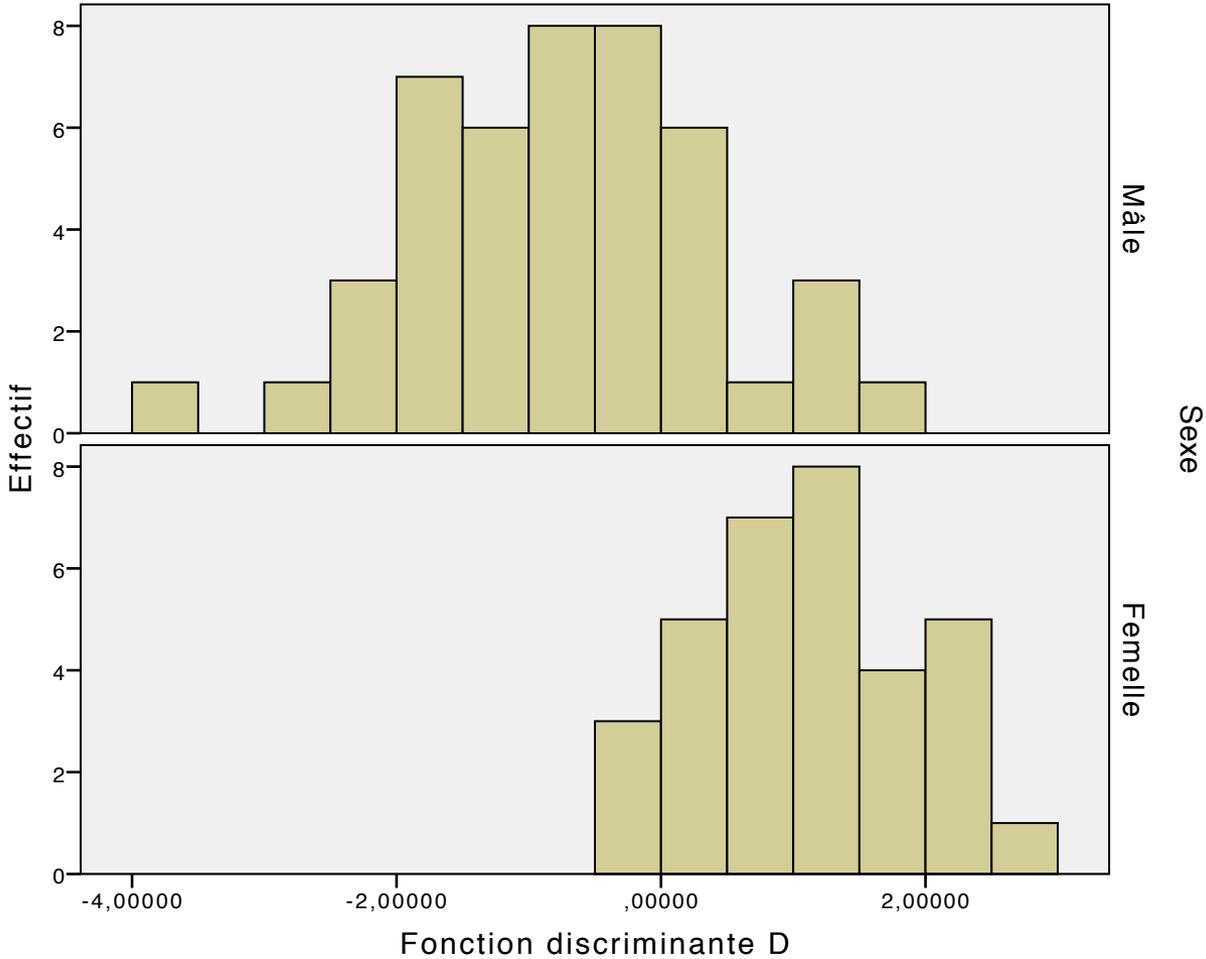
	Fonction
	1
Sexe	
Mâle	-,798
Femelle	1,088

Valeurs propres

Fonction	Valeur propre	% de la variance	% cumulé	Corrélation canonique
1	,890 <sup>a</sup>	100,0	100,0	,686

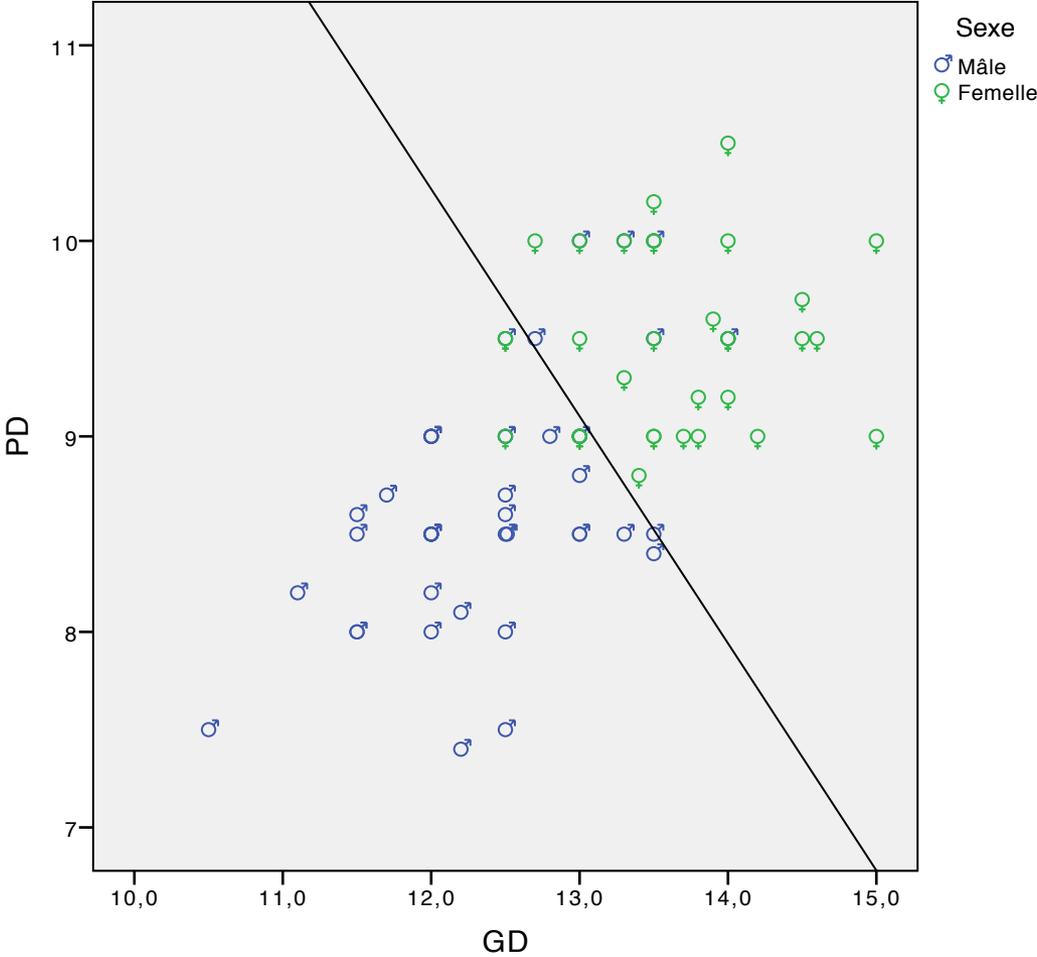
# RÉSULTATS

## Poussins de circaète



# RÉSULTATS

## Poussins de circaète



# RÉSULTATS

## Poussins de circaète

Résultats du classement<sup>a</sup>

			Classe(s) d'affectation prévue (s)		Total
			Mâle	Femelle	
Original	Effectif	Sexe			
		Mâle	39	6	45
		Femelle	6	27	33
		Observations non classées	100	96	196
	%	Mâle	86,7	13,3	100,0
		Femelle	18,2	81,8	100,0
		Observations non classées	51,0	49,0	100,0

a. 84,6% des observations originales classées correctement.

# RÉSULTATS

## Poussins de circaète

Diagnostic des observations

Nombre d'observations	Groupe effectif	Plus grand groupe					Deuxième plus grand groupe		
		Groupe prévu	P(D>d   G=g)		P(G=g   D=d)	Carré de la distance de Mahalanobis au barycentre	Groupe	P(G=g   D=d)	Carré de la distance de Mahalanobis au barycentre
			p	ddl					
Original 1	1	1	,514	1	,633	,426	2	,367	1,519
2	non classées	2	,862	1	,810	,030	1	,190	2,927
3	non classées	1	,874	1	,888	,025	2	,112	4,175
4	1	1	,303	1	,976	1,060	2	,024	8,496
5	non classées	1	,514	1	,633	,426	2	,367	1,519
6	non classées	1	,392	1	,541	,732	2	,459	1,060
7	non classées	2	,433	1	,963	,615	1	,037	7,127
8	non classées	2	,601	1	,688	,274	1	,312	1,856
9	non classées	1	,338	1	,973	,919	2	,027	8,088
11	non classées	2	,084	1	,994	2,982	1	,006	13,048

# CONCLUSION

Les fonctions discriminantes trouvées permettent de déterminer le sexe

- de tous les Autours poussins âgés d'au moins 3 semaines,
- de 85% des poussins de Circaète
- et de 89% des Buses âgés d'environ un mois, et de 95% des Buses volantes.



Les paramètres les plus utiles pour la discrimination sont les diamètres des tarse et la longueur du bec.



# UTILISATION

Le calcul des fonctions discriminantes les plus performantes, pour un échantillon de poussins dont le sexe n'était pas connu par analyse moléculaire, permet :

- de confirmer que la sexe-ratio est bien de l'ordre de 60% de femelles pour les poussins de Buse que nous étudions.
- d'évaluer à 51% (environ) la sexe-ratio des jeunes Circaètes, (sur 400 juvéniles qui ont été mesurés), alors qu'elle était de 56% dans l'échantillon de 106 individus utilisés pour l'A.D.

# UTILISATION

Cette analyse nous permettra d'aborder à nouveau des questions:

- Certaines Buses sont sédentaires, d'autres erratiques ou migratrices partielles. Nous savons que cela dépend de leur âge, et de leur statut (territorial, reproducteur ou non). **Mais peut-être aussi du sexe ?**
- Certains Circaètes restent en Afrique lors de leur deuxième printemps, d'autres reviennent en France. **Cela dépend-il de leur sexe ?**
- La longévité, la taille du territoire défendu **dépendent-ils du sexe ?**

L'échelle croissante du dimorphisme biométrique sexuel correspond au classement indiqué par Frochot et Newton :

**« The faster the prey, the greater the degree of dimorphism in the raptor »**

*(Exemple de grammaire à placer dans tous les bons manuels d'anglais !)*



Ainsi, dans les espèces que nous avons étudiées :  
**Circaète < Buse < Autour.**

