

# *STA 201 Statistique Multivariée Approfondie*

- Inscriptions et agréments (master actuariat EFAB, MR085)
- Programme
- Modalités de contrôle
- SAS <https://depotstat.cnam.fr/>
- Informations  
<http://cedric.cnam.fr/~saporta/Cours.html>

# Programme



5 octobre	Présentation du cours. Vecteurs gaussiens, domaines de confiance et de tolérance
12 octobre	Exercices ( <b>E.Jakobowicz, Addinsoft</b> )
19 octobre	Simulation, bootstrap
26 octobre	Sélection de variables en ACP ( <b>P.L.Gonzalez, CNAM</b> )
2 novembre	Traitement de la multicollinéarité en régression (ridge, PLS, RCP) 1
9 novembre	Traitement de la multicollinéarité en régression (ridge, PLS, RCP) 2
16 novembre	Codages, recodages et régression monotone
23 novembre	Analyse conjointe
30 novembre	Tableaux multiples et données évolutives ( <b>N.Niang, CNAM</b> )
7 décembre	Approche multi-tableaux ( <b>H.Abdi, UT Dallas</b> )
14 décembre	Applications de la régression PLS ( <b>H.Abdi, UT Dallas</b> )

4 janvier	Régression robuste et non-paramétrique
11 janvier	Multidimensional scaling
18 janvier	Modèles à variables latentes (analyse factorielles, classes latentes)
25 janvier	Equations structurelles 1 ( <b>E.Jakobowicz, Addinsoft</b> )
1 février	Equations structurelles 2 ( <b>E.Jakobowicz, Addinsoft</b> )
29 février	Analyse factorielle discriminante et fonction de Fisher
7 mars	Discrimination et scoring sur variables qualitatives
14 mars	Discrimination probabiliste sous hypothèses de normalité
21 mars	Mesures de performance
28 mars	Régression logistique 1
4 avril	Régression logistique 2
11 avril	Etudes de cas ( <b>N.Niang, CNAM</b> )
2 mai	Estimation de densité, discriminante non-paramétrique
9 mai	Machines à vecteurs de support (SVM)
16 mai	Introduction à la théorie de l'apprentissage
23 mai	<i>Défini ultérieurement</i>
30 mai	Inférence sur les valeurs propres en ACP, AFC
6 juin	Données fonctionnelles 1 ( <b>C.Preda, Polytech'Lille</b> )
13 juin	Données fonctionnelles 2 ( <b>C.Preda, Polytech'Lille</b> )
20 juin	<i>Défini ultérieurement</i>

# Bibliographie



- M.BARDOS : *Analyse discriminante* (Dunod, 2001)
- B.ESCOFIER, J.PAGES : *Analyses factorielles simples et multiples*, 4<sup>ème</sup> édition (Dunod, 2008)
- T.HASTIE, J.FRIEDMAN, R.TIBSHIRANI : *The Elements of Statistical Learning*, 2<sup>ème</sup> édition (Springer, 2009)
- L.LEBART, M.PIROU, A.MORINEAU : *Statistique exploratoire multidimensionnelle*, 4<sup>ème</sup> édition (Dunod, 2006)
- J.P.NAKACHE, J.CONFAIS : *Statistique explicative appliquée* (Technip, 2003)
- G.SAPORTA : *Probabilités, analyse des données, statistique*, 3<sup>ème</sup> édition (Technip, 2011)
- S.TUFFERY : *Data mining et statistique décisionnelle*, 3<sup>ème</sup> édition (Technip, 2010)

# *Eléments d'analyse multivariée*

*Gilbert Saporta*

*Conservatoire National des Arts et Métiers,  
Paris*

*[gilbert.saporta@cnam.fr](mailto:gilbert.saporta@cnam.fr)*

*<http://cedric.cnam.fr/~saporta>*

*octobre 2011*

# *I. Généralités sur les vecteurs aléatoires*

- **I.1 Fonction de répartition et densité**

$x \in \mathbb{R}^p$

$$F(x_1, \dots, x_p) = P(X_1 < x_1 \cap \dots \cap X_p < x_p)$$

$$f(x_1, \dots, x_p) = \frac{\partial^p F(x_1, \dots, x_p)}{\partial x_1 \dots \partial x_p}$$

- Changement de variable

$$\mathbf{Y} = \varphi(\mathbf{X})$$

$$Y_i = \varphi_i(X_1, \dots, X_p)$$

- Densité

$$g(\mathbf{y}) = \frac{f(\varphi^{-1}(\mathbf{y}))}{|\det \mathbf{J}|}$$

- J jacobien

$$\mathbf{J} = \begin{vmatrix} \frac{\partial y_1}{\partial x_1} & \dots & \frac{\partial y_1}{\partial x_p} \\ \dots & & \dots \\ \frac{\partial y_p}{\partial x_1} & \dots & \frac{\partial y_p}{\partial x_p} \end{vmatrix} = \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \dots & \frac{\partial x_1}{\partial y_p} \\ \dots & & \dots \\ \frac{\partial x_p}{\partial y_1} & \dots & \frac{\partial x_p}{\partial y_p} \end{vmatrix}^{-1}$$

- **I.2 Fonction caractéristique** (transformée de Fourier de la densité)

$$\varphi_{\mathbf{X}}(\mathbf{a}) = E(\exp i\mathbf{a}'\mathbf{x}) = E(\exp i(\sum_{j=1}^p a_j X_j))$$

- Condition nécessaire et suffisante d'indépendance

$$\varphi_{\mathbf{X}}(\mathbf{a}) = \prod_{j=1}^p E(\exp i a_j X_j) = \prod_{j=1}^p \varphi_j(a_j)$$

- Théorème de Cramer-Wold:
  - La loi de  $X$  est déterminée par celle des combinaisons linéaires

$$Y = \mathbf{a}'\mathbf{x} = \sum_{j=1}^p a_j X_j$$

$$\varphi_Y(t) = E(\exp(itY)) = E(\exp(it\mathbf{a}'\mathbf{x}))$$

$$\varphi_Y(1) = \varphi_{\mathbf{X}}(\mathbf{a})$$

## ■ I.3 Espérance et matrice de variance

$$\boldsymbol{\mu} = \begin{pmatrix} E(X_1) \\ \cdot \\ \cdot \\ E(X_p) \end{pmatrix} \quad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & & \\ \dots & & & \\ & & & \sigma_p^2 \end{pmatrix} = E(\mathbf{xx}') - \boldsymbol{\mu}\boldsymbol{\mu}'$$

- Transformation linéaire  $\mathbf{y} = \mathbf{A}\mathbf{x}$

$$\boldsymbol{\mu}_y = \mathbf{A}\boldsymbol{\mu}_x \quad \boldsymbol{\Sigma}_y = \mathbf{A}\boldsymbol{\Sigma}_x\mathbf{A}'$$

- Cas particulier  $y = \mathbf{a}'\mathbf{x}$

$$V(y) = \mathbf{a}'\boldsymbol{\Sigma}\mathbf{a}$$

- CNS pour être une matrice de variance:
  - $\Sigma$  symétrique positive
- $\Sigma$  a une racine carrée:  $\Sigma^{1/2}$

## ■ I.4 Transformation de Mahalanobis:

$$\mathbf{y} = \Sigma^{-1/2}(\mathbf{x} - \boldsymbol{\mu})$$

- Composantes de  $\mathbf{Y}$  centrées, réduites, non-corrélées
- Distance de Mahalanobis au centre:

$$D^2 = (\mathbf{x} - \boldsymbol{\mu})' \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}) = \mathbf{y}' \mathbf{y}$$

$$E(D^2) = p \quad \text{car } D^2 = \sum_{j=1}^p Y_j^2$$

- Pour toute matrice orthogonale  $\mathbf{P}$  ( $\mathbf{P}' = \mathbf{P}^{-1}$ ),  $\mathbf{P}\mathbf{y}$  a aussi ses composantes centrées, réduites et non-corrélées.

# *1.5 Composantes principales*

- Combinaisons linéaires de variance maximale :  $V(y) = \mathbf{a}'\Sigma\mathbf{a}$
- $\mathbf{a}$  vecteur propre de  $\Sigma$

# *II Loi normale à $p$ dimensions*



## ■ **II.1 Définition**

- $X$  suit une  $N_p$  si et seulement si toute combinaison linéaire des composantes suit une loi normale
- $N_p(\mu; \Sigma)$
- Les composantes de  $X$  sont indépendantes si et seulement si  $\Sigma$  est diagonale

## ■ II.2 Densité

$$f(x_1, \dots, x_p) = \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})' \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})\right)$$

- Démonstration: changement de variable:  $\mathbf{y}$  est  $N_p(0; I)$

$$\mathbf{y} = \Sigma^{-1/2} (\mathbf{x} - \boldsymbol{\mu})$$

$$g(y) = \prod_{j=1}^p \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} y_j^2\right) = \frac{1}{(2\pi)^{p/2}} \exp\left(-\frac{1}{2} \sum_{j=1}^p y_j^2\right)$$

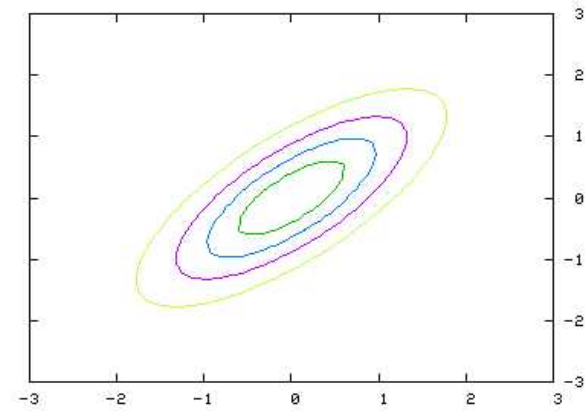
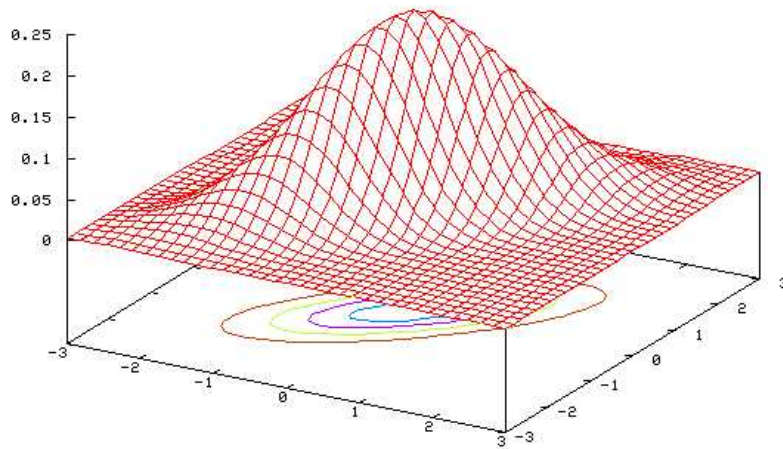
puis jacobien et retour à  $\mathbf{x}$

- II.3 loi de  $D^2$  et surfaces d'isodensité

$$D^2 = (\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) = \mathbf{y}' \mathbf{y} = \sum_{j=1}^p Y_j^2$$

- $D^2$  suit une loi du khi-deux à  $p$  degrés de liberté
- $f(\mathbf{x}) = c \Rightarrow D^2 = k$  : ellipsoïde

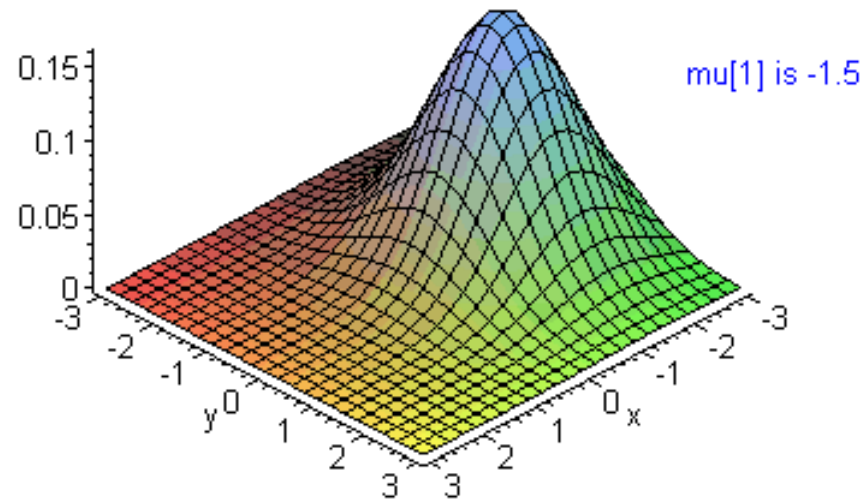
- Loi normale à deux dimensions



Source: <http://azzalini.stat.unipd.it/SN/plot-SN2.html>

- Loi normale à deux dimensions

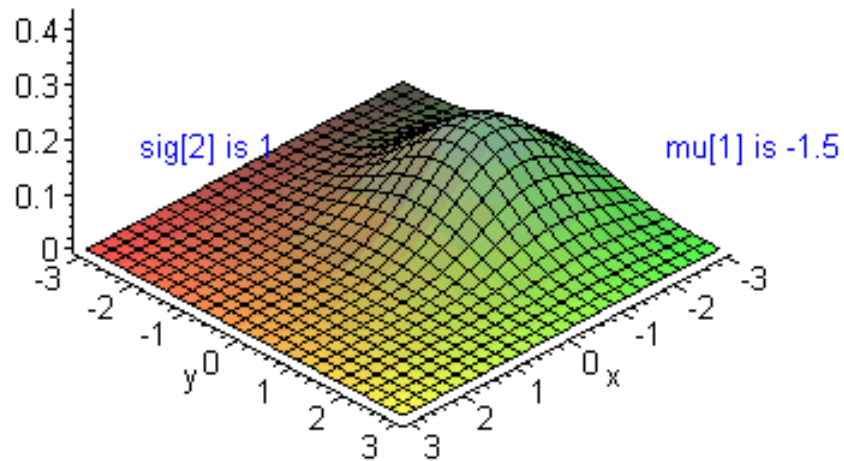
mu[1] is changing!



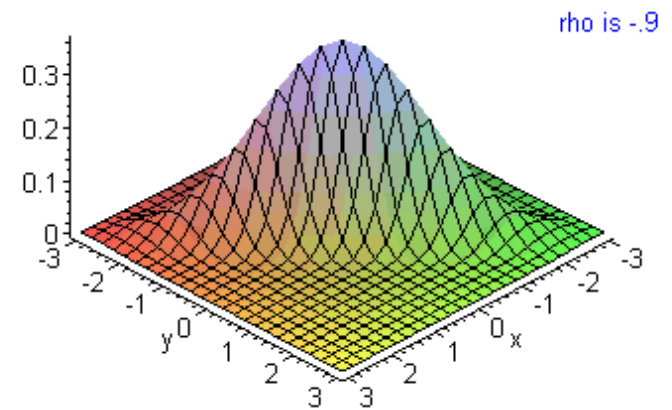
- Source : <http://www2.kenyon.edu/people/hartlaub/MellonProject/Bivariate2.html>

# ■ Loi normale à deux dimensions

sigma[2] is changing, too!



rho across (-1,1)



# III Quelques problèmes d'inférence

## ■ III.1 Centre de gravité d'un nuage de points

- n observations iid

$$\begin{pmatrix} x_1^1 & \dots & x_1^p \\ \vdots & \ddots & \vdots \\ x_n^1 & \dots & x_n^p \end{pmatrix}$$

$$\mathbf{g} = (\bar{x}^1; \bar{x}^2; \dots; \bar{x}^p)$$

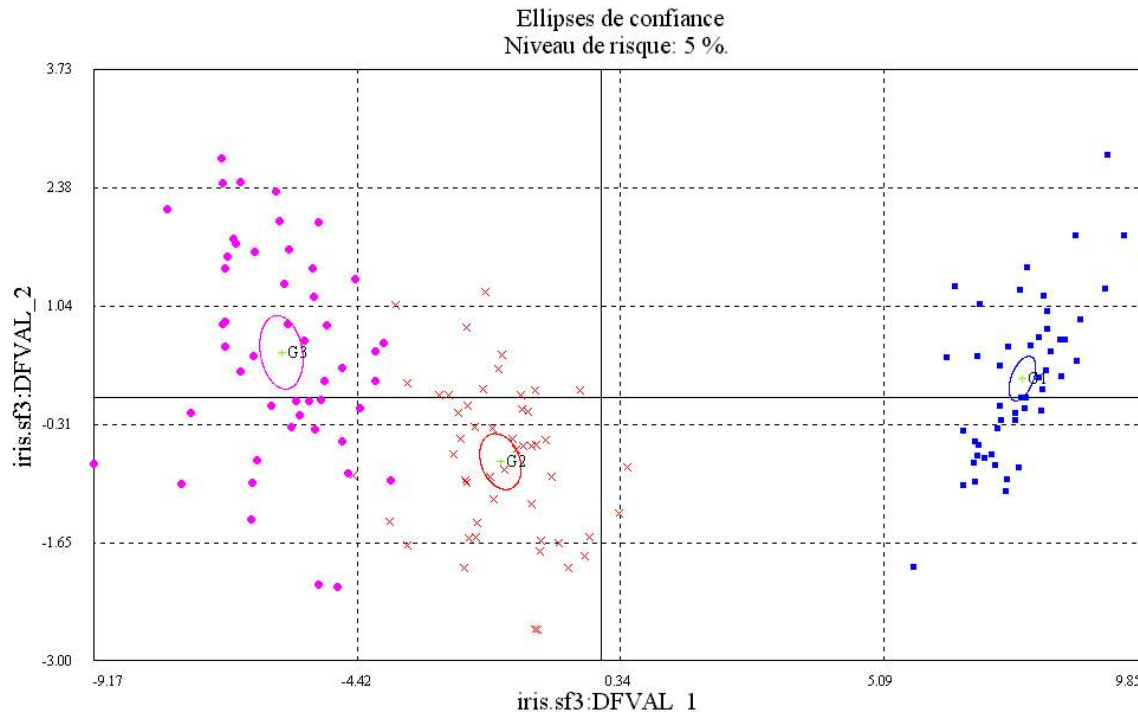
$$E(\mathbf{g}) = \boldsymbol{\mu} \quad \boldsymbol{\Sigma}_{\mathbf{g}} = \frac{\boldsymbol{\Sigma}}{n}$$

- Théorème central-limite
  - Pour  $n$  grand  $\mathbf{g}$  suit approximativement une loi  $N_p(\boldsymbol{\mu}; \boldsymbol{\Sigma}/n)$
- Ellipsoïde de confiance:
  - si  $\boldsymbol{\Sigma}$  est connu

$$(\mathbf{g} - \boldsymbol{\mu})' \left( \frac{\boldsymbol{\Sigma}}{n} \right)^{-1} (\mathbf{g} - \boldsymbol{\mu}) \quad \text{suit une loi } \chi_p^2$$

- si  $\Sigma$  est inconnu: estimation par  $V$  matrice de variance-covariance de l'échantillon ( $n$  en dénominateur)
- La loi du khi-deux doit être remplacée par une  $F$ :

$$(\mathbf{g} - \boldsymbol{\mu})' \mathbf{V}^{-1} (\mathbf{g} - \boldsymbol{\mu}) = \frac{p}{n - p} F(p; n - p)$$



- **III. 2 Ellipsoïde de tolérance: où se trouve une nouvelle observation?**

$$(\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) = k$$

où  $k$  est le fractile  $1-\alpha$  d'un  $\chi_p^2$

- Si  $\boldsymbol{\mu}$  est estimé par  $\mathbf{g}$ ,

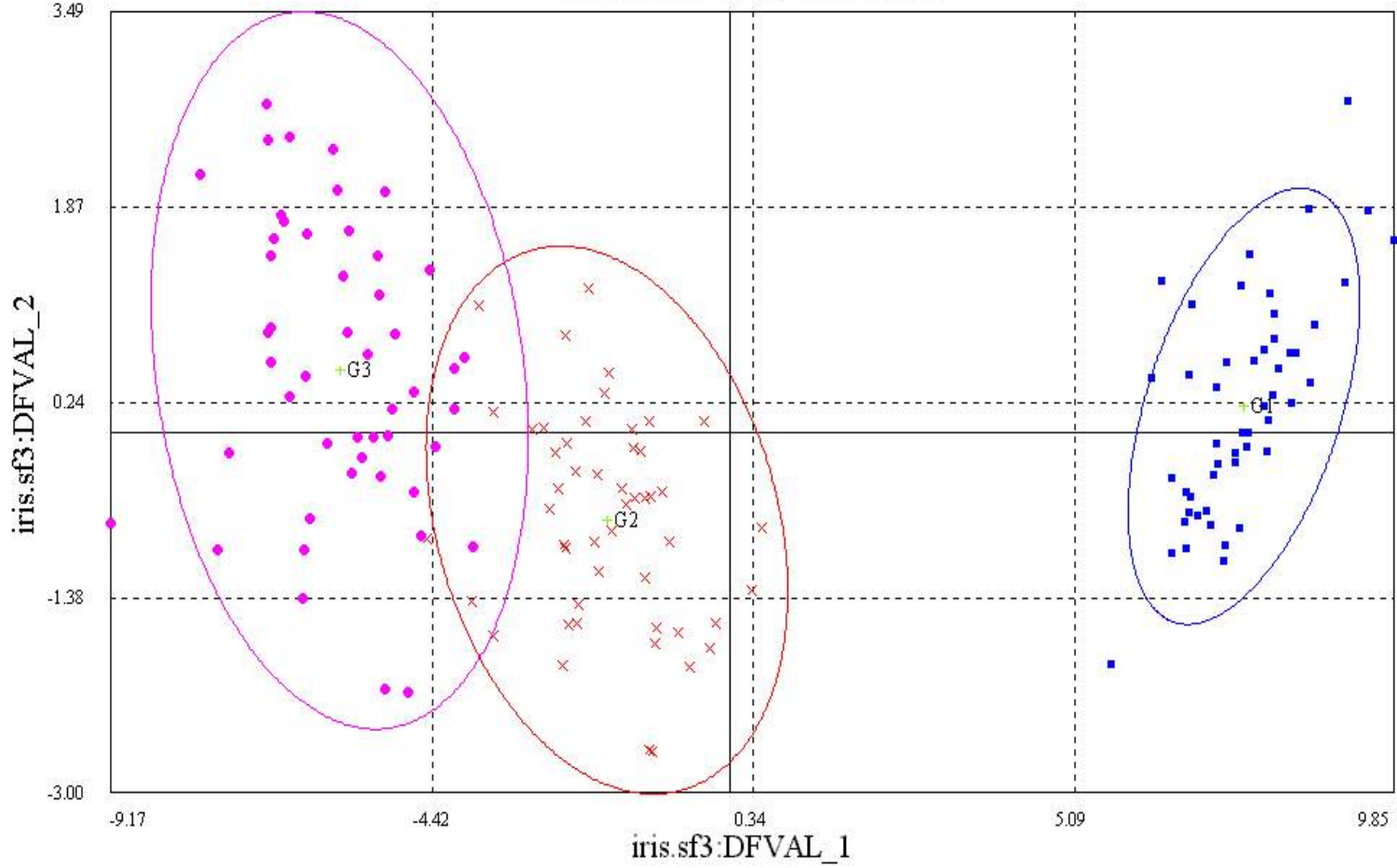
$\mathbf{x}-\mathbf{g}$  suit une

$$N_p \left( \mathbf{0}; \boldsymbol{\Sigma} \left( 1 + \frac{1}{n} \right) \right)$$

Si  $\boldsymbol{\Sigma}$  estimé par la matrice de variance du nuage  $\mathbf{V}$ :

$$(\mathbf{x} - \mathbf{g})' \mathbf{V}^{-1} (\mathbf{x} - \mathbf{g}) = \frac{(n-1)p}{n-p} \frac{n+1}{n} F(p; n-p)$$

Ellipses de tolérance  
contenant 95 % des observations



# III.3 Distance de Mahalanobis



Distance au sens de la métrique  $W^{-1}$ .

$$D_p^2 = (g_1 - g_2)' W^{-1} (g_1 - g_2)$$

1. pour  $p=1$  : 
$$\frac{n_1 n_2}{n_1 + n_2} \left( \frac{\bar{x}_1 - \bar{x}_2}{\hat{\sigma}} \right)^2 = \frac{n_1 n_2}{n_1 + n_2} D_1^2 \sim F(1; n_1 + n_2 - 2)$$

2.  $p$  quelconque :

$$D_p^2 = (g_1 - g_2)' W^{-1} (g_1 - g_2)$$

$$D_p^2 = (g_1 - g_2)' W^{-1/2} \underbrace{W^{-1/2} (g_1 - g_2)}_{W^{-1/2} X}$$

- *Standardisation de chaque composante  $x_j$*
- *Décorrélation...*

Le  $\Delta^2$  théorique :  $\Delta_p^2 = (\underline{\mu}_1 - \underline{\mu}_2)' \Sigma^{-1} (\underline{\mu}_1 - \underline{\mu}_2)$

2 populations  $N_p (\underline{\mu}_1, \Sigma)$  et  $N_p (\underline{\mu}_2, \Sigma)$

$D_p^2$  estimation (biaisée) de  $\Delta_p^2$

$$W = \frac{n_1 V_1 + n_2 V_2}{n - 2} = \hat{\Sigma}$$

$$D_p^2 = (\underline{g}_1 - \underline{g}_2)' W^{-1} (\underline{g}_1 - \underline{g}_2)$$

# Test de comparaison de deux moyennes multivariées

$$E(D_p^2) = \frac{n-2}{n-p-1} \left( \Delta_p^2 + \frac{pn}{n_1 n_2} \right)$$

$$\text{Si } \Delta^2 = 0 \quad \underline{\mu}_1 = \underline{\mu}_2$$

$$\frac{n_1 n_2}{n} \frac{n-p-1}{p(n-2)} D_p^2 \sim F(p; n-p-1)$$

# Forme de Hotelling

$$\frac{n(n-2)}{n^2} (\mathbf{g}_1 - \mathbf{g}_2)' \mathbf{W}^{-1} (\mathbf{g}_1 - \mathbf{g}_2) = T_p^2(n-2)$$

$$T_p^2(n) = \frac{np}{n-p+1} F(p; n-p+1)$$

## 2 groupes parmi k

- Théoriques :

$$\Delta_p^2 = (\mu_i - \mu_j)' \Sigma^{-1} (\mu_i - \mu_j)$$

- Estimées :

$$D_p^2 = (\underline{g}_i - \underline{g}_j)' \left( \frac{n}{n-k} W \right)^{-1} (\underline{g}_i - \underline{g}_j)$$

$$\text{Si } \Delta^2 = 0$$

$$\frac{n_i n_j}{n_i + n_j} \cdot \frac{n-k-p+1}{(n-k)p} \cdot D_p^2 = F(p; n-k-p+1)$$