STA 201 Statistique Multivariée Approfondie

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

Bibliographie

- M.BARDOS : *Analyse discriminante* (Dunod, 2001)
- T.HASTIE, J.FRIEDMAN, R.TIBSHIRANI : *The Elements of Statistical Learning*, 2ème édition (Springer, 2009)
  http://www.stanford.edu/~hastie/local.ftp/Springer/ESLII_print10.pdf
- L.LEBART, M.PIRON, A.MORINEAU : *Statistique exploratoire multidimensionnelle*, 4ème édition (Dunod, 2006)
- J.P.NAKACHE, J.CONFAIS : *Statistique explicative appliquée* (Technip, 2003)
- G.SAPORTA : *Probabilités, analyse des données, statistique*, 3ème édition (Technip, 2011)
- S.TUFFÉRY : *Data mining et statistique décisionnelle*, 3ème édition (Technip, 2010)

Eléments d'analyse multivariée

Gilbert Saporta

Conservatoire National des Arts et Métiers, Paris

gilbert.saporta@cnam.fr

http://cedric.cnam.fr/~saporta

octobre 2013

I.Généralités sur les vecteurs aléatoires

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

  $x \in R^p$

$$F(x_1,...,x_p) = P(X_1 < x_1 \bigcap ... \bigcap X_p < x_p)$$

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

- Changement de variable

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

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

- Densité

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

- J jacobien

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

- **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** (1936):
  - La loi de X est déterminée de façon unique par celle des combinaisons linéaires des composantes

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

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

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

Harald Cramer
1893-1985

Herman Wold
1908-1992

I.3 Espérance et matrice de variance

$$\boldsymbol{\mu} = \begin{pmatrix} E(X_1) \\ . \\ . \\ E(X_p) \end{pmatrix} \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \\ ... & & \\ .... & & \sigma_p^2 \end{pmatrix} = E(\mathbf{x}\mathbf{x}') - \boldsymbol{\mu}\boldsymbol{\mu}'$$

- Transformation linéaire **y=Ax**

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

- Cas particulier y=a'x

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

- CNS pour être une matrice de variance:
  - Σ symétrique positive

- Σ a une racine carrée: $\Sigma^{1/2}$

- **I.4 Transformation de Mahalanobis:**

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

  - Composantes de Y centrées, réduites, non-corrélées

- Distance de Mahalanobis au centre:

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

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

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

Prasanta Chandra Mahalanobis

1893-1972

- Fondateur de l'Indian Statistical Institute 1931
- *On the generalized distance in statistics*, 1936
- Mahalanobis International Award in Statistics depuis 2003

I.5 Composantes principales

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

- $\mathbf{a}$ vecteur propre de $\boldsymbol{\Sigma}$

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 Σ est diagonale

- **II.2 Densité**

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

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

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

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

puis jacobien et retour à **x**

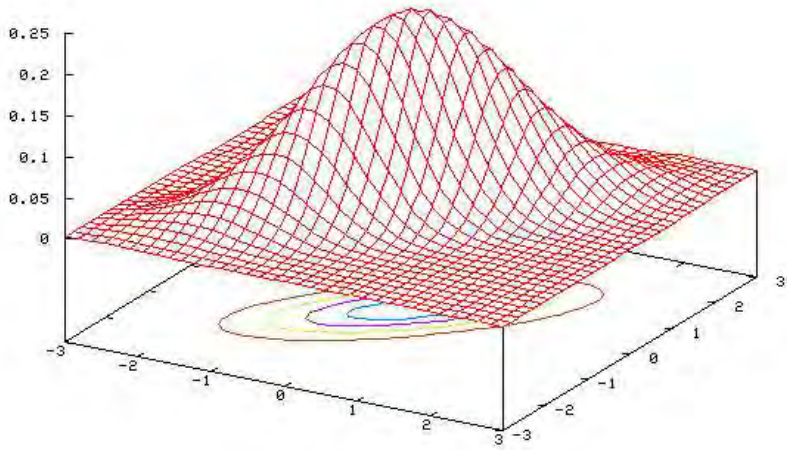
- II.3 loi de D^2 et surfaces d'isodensité

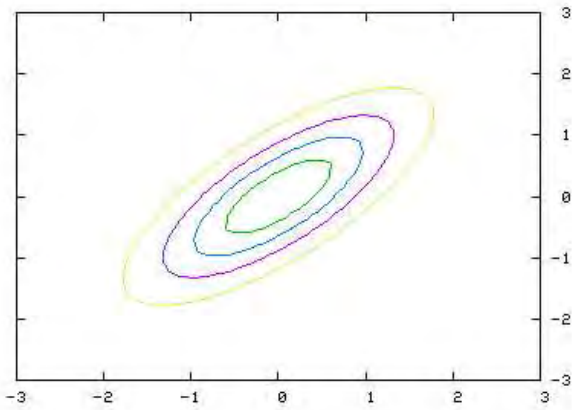
$$D^2 = (\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) = \mathbf{y}'\mathbf{y} = \sum_{j=1}^{p} Y_j^2$$

- D^2 suit une loi du khi-deux à p degrés de liberté

- $f(x)=c \Rightarrow D^2=k$: ellipsoïde

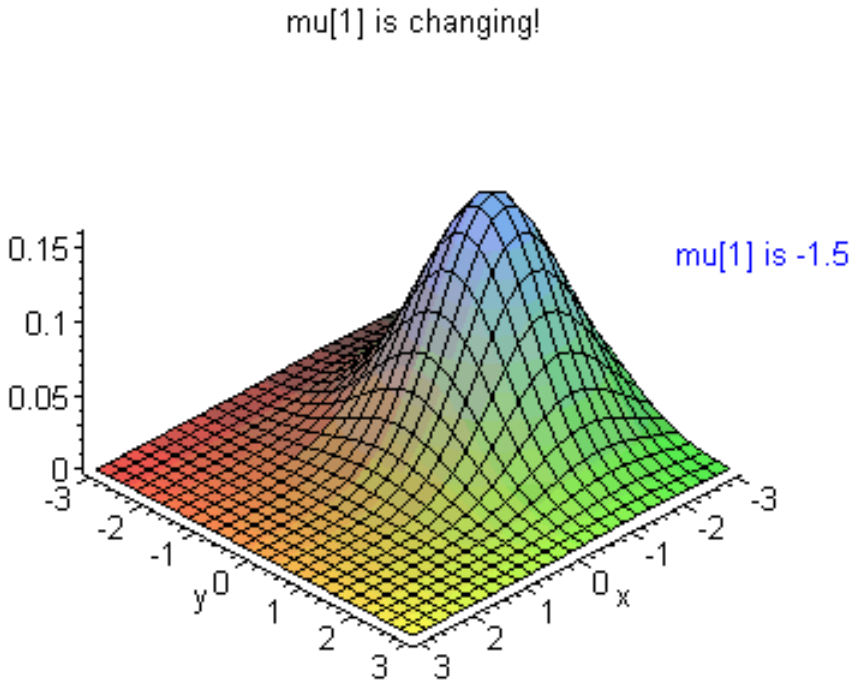
Loi normale à deux dimensions





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

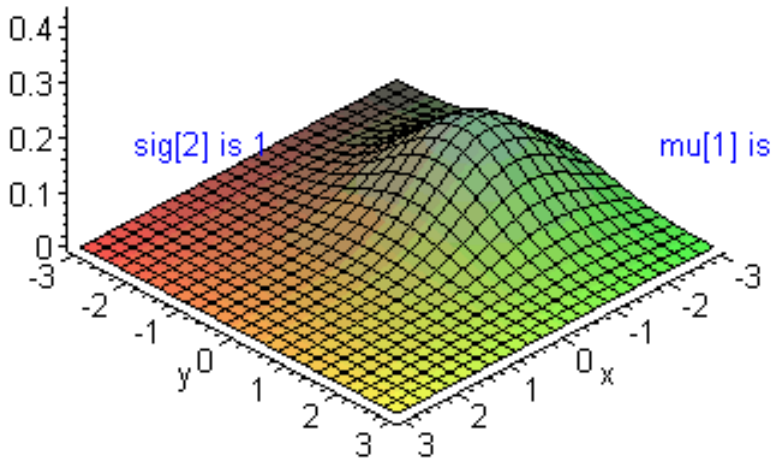
Loi normale à deux dimensions



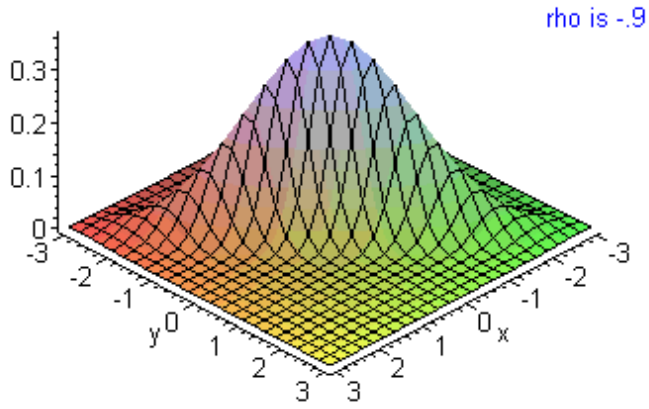
- Source : http://www2.kenyon.edu/people/hartlaub/MellonProject/Bivariate2.html

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

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

$$\mathbf{g} = (\overline{x}^1; \overline{x}^2; ....; \overline{x}^p)$$

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

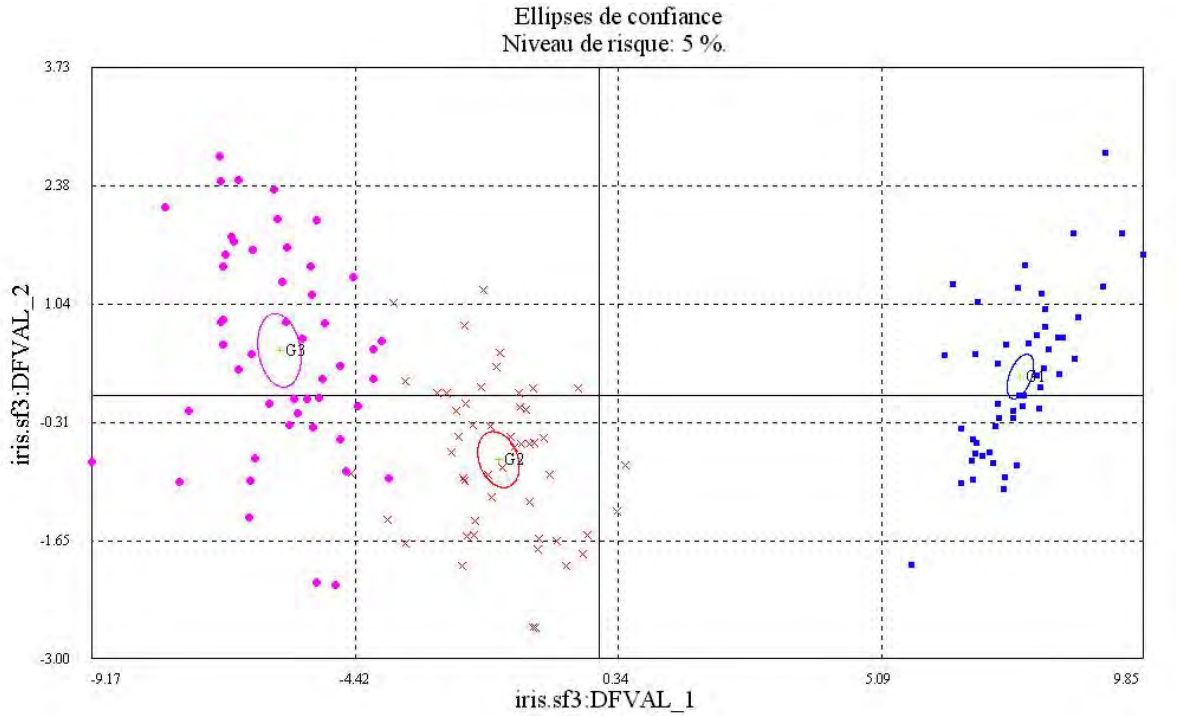
- Théorème central-limite
  - Pour n grand g suit approximativement une loi $N_p(\mu; \Sigma/n)$

- Ellipsoïde de confiance:
  - si Σ est connu

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

- si Σ 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?

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

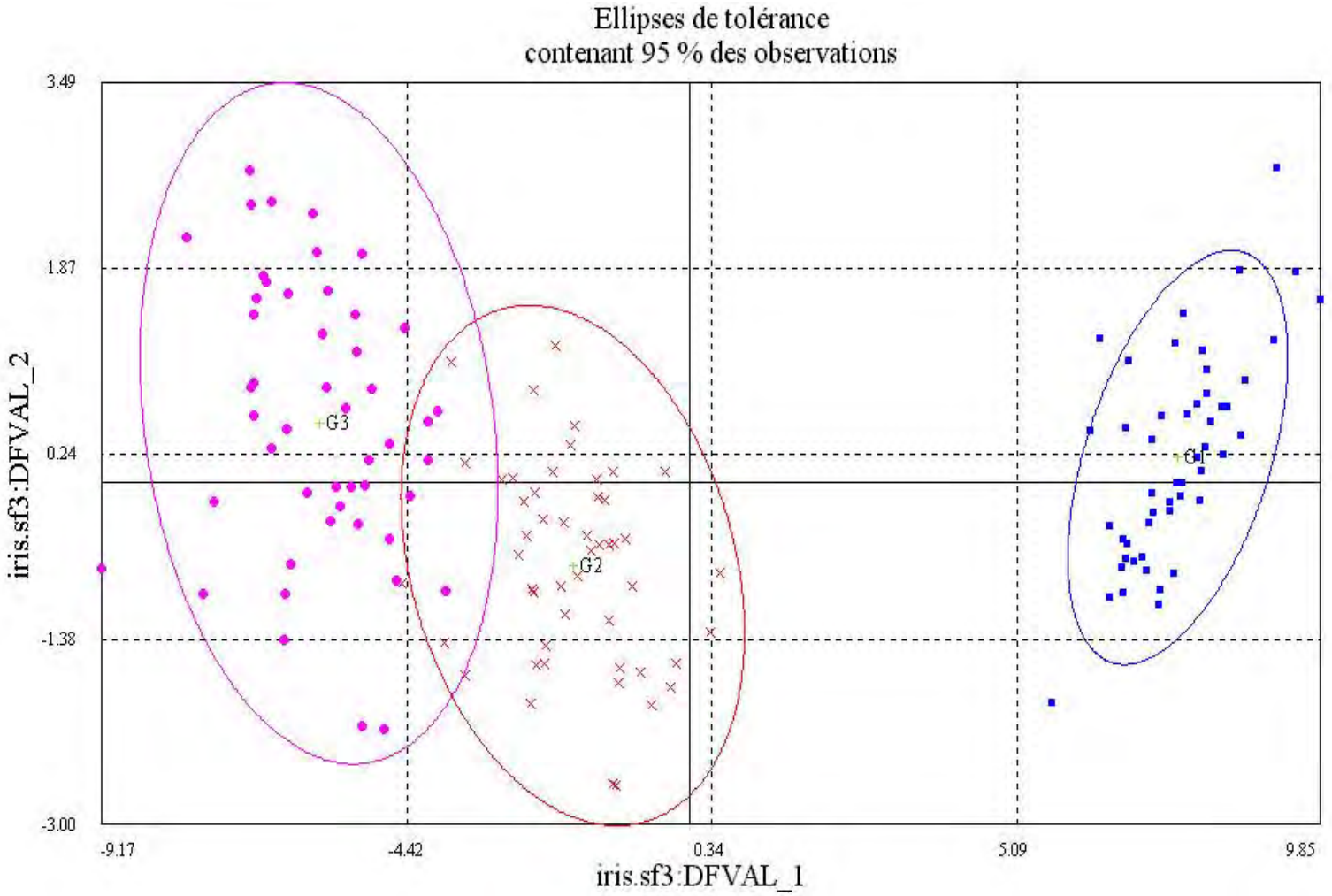
où k est le fractile 1-α d'un χ^2_p

- Si μ est estimé par g,

  x-g suit une $N_p\left(\mathbf{0};\boldsymbol{\Sigma}\left(1+\frac{1}{n}\right)\right)$

Si Σ estimé par la matrice de variance du nuage 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
3.49
1.87
0.24
-1.38
-3.00
iris.sf3:DFVAL_2
-9.17
-4.42
0.34
5.09
9.85
iris.sf3:DFVAL_1
G3
G2
G1

III.3 Distance de Mahalanobis

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

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

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

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

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

D_p^2 estimation (biaisée) de Δ_p^2

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

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

Test de comparaison de deux moyennes multivariées

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

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

$$\frac{n_1 n_2}{n}\ \frac{n-p-1}{p(n-2)}\ \mathrm{D}_\mathrm{p}^2 \sim F\left(p; n-p-1\right)$$

Forme de Hotelling

$$\frac{n(n-2)}{n^2}(\mathbf{g_1 - g_2})'\mathbf{W}^{-1}(\mathbf{g_1 - g_2}) = T_p^2(n-2)$$

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

2 groupes parmi k

- Théoriques :

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

- Estimées :

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

$$Si \quad \Delta^2 = 0$$

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