

## **Statistiques et informatique**

### **UV PSY33A1**

## **Présentation du cours 2002/2003**

### **Organisation matérielle**

Cours magistral: 13 heures. Statistiques et Informatique

A peu près 8 heures de statistiques et  
5 heures d'informatique  
mardi 9h15-10h15 - Amphi 3

Travaux dirigés:

3 groupes de TD de statistiques.

lundi 13h45-14h45 - A009 - Gr 2

lundi 14h45-15h45 - Gr 3

mercredi 9h15-10h15 - Gr 1

5 sous-groupes de TD en Informatique.

salles A204 et A206, 2 h. par quinzaine

1 sous-gr - jeudi 8h15-10h20 - sem. B

2 sous-gr - Jeudi 8h15-10h20

2 sous-gr - vendredi 8h15-10h20

Monitorat informatique

Contrôle des connaissances: (contrôle continu)

70 % Examen écrit (3 heures)

30 % Note de TD

## Bibliographie

- B. Cadet Méthodes statistiques en psychologie. P.U. de Caen
- G. Mialaret. Statistiques appliquées aux sciences humaines. PUF
- B. Beaufils. Statistiques appliquées à la psychologie. Tomes 1 et 2. LEXIFAC Bréal
- N. Guéguen. Manuel de statistiques pour psychologues
- D.C. Howell. Méthodes statistiques en sciences humaines
- M. Reuchlin. Précis de Statistiques. PUF Coll. Le Psychologue.
- D. Laveault, J. Grégoire. Introduction aux théories des tests en sciences humaines. De Boeck Université
- J.P. Rossi. La méthode expérimentale en Psychologie
- P. Rateau, Méthode et statistique expérimentales en sciences humaines, Ellipses

## **Documents fournis :**

Transparents du cours de statistiques

Polycopié du cours en Informatique

Fiches de TD de statistiques et d'informatique

Documents de l'année 2001/02 disponibles sur internet

– Au format .pdf lisible par Acrobat Reader:

Transparents du CM de Stats, fiches de TD de Stats

– Au format HTML (navigateur Web) :

fiches de TD d'informatique

### *Adresses Web*

– accès en dehors des salles informatiques :

<http://geai.univ-brest.fr/~carpentier/>

<http://infolettres.univ-brest.fr/~carpentier/>

– accès depuis les salles informatiques :

<http://letsamba.univ-brest.fr/~carpentier/>

## **Contenu**

### **Statistiques :**

Echantillonnage

Estimation de paramètres

Tests statistiques: comparaison de moyennes, tests non paramétriques

### **Informatique :**

Comment l'informatique est-elle perçue?

Homme et ordinateur: quelques idées sur l'interface utilisateur.

Hypertextes et hypermedias.

Ordinateur et apprentissage.

## Echantillonnage

### Echantillonnage - cas d'une moyenne

$\mu$  : moyenne sur la population

$\sigma^2$  : variance sur la population

Distribution d'échantillonnage de  $\bar{X}$ , moyenne observée sur un échantillon de taille  $n$ .

Loi normale (si  $n \geq 30$ )

Moyenne :  $Moy(\bar{X}) = \mu$

Variance, appelée erreur standard ou erreur type :

$$Var(\bar{X}) = \frac{\sigma^2}{n}$$

### Echantillonnage - cas d'une proportion

$p$  : proportion dans la population

$\sigma^2 = p(1 - p)$

Distribution d'échantillonnage de  $\bar{F}$ , proportion observée sur un échantillon de taille  $n$ .

Loi normale (si  $np(1 - p) \geq 20$ )

Moyenne :  $Moy(\bar{F}) = p$

Variance :  $Var(\bar{F}) = \frac{p(1 - p)}{n}$

## Estimation de paramètres

### Statistiques inférentielles

Raisonnement de type inductif: à partir de *conséquences* (ce qui est observé sur un (ou des) échantillons de taille  $n$ ), remonter aux *causes* les plus probables (valeurs des paramètres dans la population).

### Estimation ponctuelle de paramètres

Population:  $\mu, \sigma^2$  inconnus.

Echantillon de taille  $n$ :  $\bar{x}, s^2$  observés.

Estimation de  $\mu$ :  $\hat{\mu} = \bar{x}$

Estimation de  $\sigma^2$ :  $\hat{\sigma}^2 = s_c^2 = \frac{n}{n-1}s^2$

$s_c^2$  est appelée **variance corrigée**.

## Intervalle de Confiance

Population nombreuse.

Variable: score obtenu à un test

Echantillon:  $n = 100$ ,  $\bar{x}_{obs} = 44$ ,  $s^2 = 142.6$ ,  $s_c^2 = 144$

Distribution d'échantillonnage:

Loi normale

Moyenne:  $\mu$

Variance estimée:  $\frac{144}{100} = 1.44$ . Ecart type: 1.2

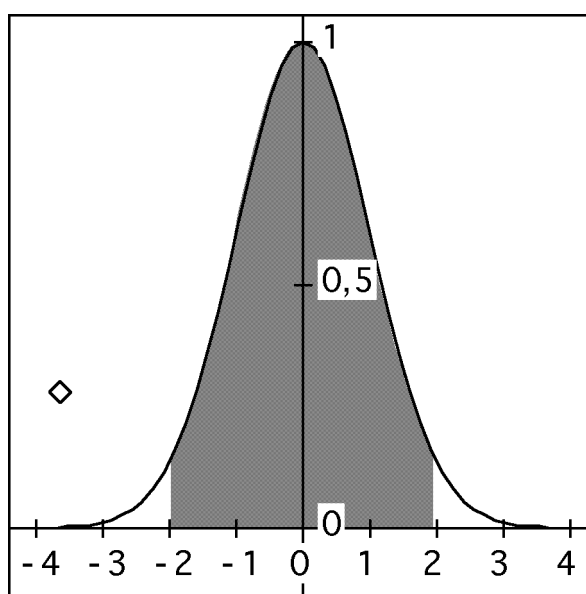
Rappel: Pour une loi normale de paramètres  $\mu$  et  $\sigma$ , l'intervalle  $[\mu - 1.96\sigma, \mu + 1.96\sigma]$  rassemble 95% des observations.

Avec 95% de degré de confiance, on a :

$$\mu - 1.96 \times 1.2 \leq 44 \leq \mu + 1.96 \times 1.2$$

ou encore

$$44 - 1.96 \times 1.2 \leq \mu \leq 44 + 1.96 \times 1.2$$



## **Intervalle de confiance pour une moyenne Cas des grands échantillons ( $n \geq 30$ )**

Problème:  $\mu$  inconnu. Estimer  $\mu$  avec un degré de confiance  $1 - \alpha$  connaissant  $\bar{x}_{obs}$ ,  $s^2$ ,  $n$ .

$$\text{Erreur type: } E^2 = \frac{s_c^2}{n}$$

On a, avec le degré de confiance  $1 - \alpha$ :

$$\bar{x}_{obs} - z_\alpha E \leq \mu \leq \bar{x}_{obs} + z_\alpha E$$

$z_\alpha$ : valeur lue dans la table de la loi normale centrée réduite, telle que:  $P(-z_\alpha \leq Z \leq z_\alpha) = 1 - \alpha$ .

## **Intervalle de confiance pour une moyenne Cas des petits échantillons ( $n < 30$ )**

Avec les mêmes notations, on a avec le degré de confiance  $1 - \alpha$ :

$$\bar{x}_{obs} - t_\alpha E \leq \mu \leq \bar{x}_{obs} + t_\alpha E$$

$t_\alpha$ : valeur lue dans la table de la **loi de Student** pour le seuil  $\alpha$  et le *nombre de degrés de liberté*  $ddl = n - 1$ .

## **Intervalle de confiance pour une proportion Cas des grands échantillons ( $np(1 - p) \geq 20$ )**

Problème:  $p$  inconnu. Estimer  $p$  avec un degré de confiance  $1 - \alpha$  connaissant  $f_{obs}$ ,  $n$ .

$$\text{Erreur type: } E^2 = \frac{f_{obs}(1 - f_{obs})}{n}$$

On a, avec le degré de confiance  $1 - \alpha$ :

$$f_{obs} - z_{\alpha}E \leq p \leq f_{obs} + z_{\alpha}E$$

$z_{\alpha}$ : valeur lue dans la table de la loi normale centrée réduite, telle que:  $P(-z_{\alpha} \leq Z \leq z_{\alpha}) = 1 - \alpha$ .



## Introduction aux tests statistiques

### Démarche générale d'un test

15 sujets soumis à un apprentissage. Deux tests l'un avant, l'autre après l'apprentissage.

Sujet	1	2	...
Avant	8	13	...
Après	11	11	...

**Problème :** L'apprentissage a-t-il un effet sur la performance ?

### Remarques :

Raisonnement en termes "d'échantillon tiré d'une population"

Variable pertinente : différence individuelle  $d_i = y_i - x_i$

### Protocole dérivé des différences individuelles

Sujet	1	2	3	4	5	6	7	8
$d_i$	3	-2	2	4	-2	1	5	3
Sujet	9	10	11	12	13	14	15	
$d_i$	4	-1	4	-1	3	3	2	

Caractéristiques de position et de dispersion :

$$\bar{d} = 1.87 ; s^2 = 5.05 ; s = 2.25 ; s_c^2 = 5.41 ; s_c = 2.33$$

## Construction d'un test statistique

Sujets observés : échantillon tiré dans une population  
 $\delta$  : moyenne des effets individuels dans la population.

### 1. Formulation des hypothèses

$H_0$  : hypothèse nulle :  $\delta = 0$

$H_1$  : hypothèse alternative :  $\delta \neq 0$

### 2. Choix d'un risque, ou seuil de signification

Par exemple :  $\alpha = 5\%$

### 3. Choix d'une statistique de test

Une statistique est une variable qui peut être évaluée sur chaque échantillon tiré, et dont la distribution théorique, sous l'hypothèse  $H_0$ , est connue.

Ici, on prend :  $T = \frac{\bar{d}}{E}$  avec  $E^2 = \frac{s_c^2}{n}$ .

Les statisticiens ont montré que, sous l'hypothèse  $H_0$ ,  $T$  suit une loi de Student à  $(n - 1)$  ddl.

### 4. Calcul des valeurs critiques (règle de décision)

Pour  $ddl = 14$  et  $\alpha = .05$ , on obtient  $t_{crit} = 2.15$ .

### 5. Calcul de la valeur observée de la statistique

Ici :  $t_{obs} = \frac{1.87}{0.60} = 3.11$

### 6. Comparer $t_{obs}$ et $t_{crit}$ . Appliquer la règle de décision

Ici :  $t_{obs} > t_{crit}$ .  $t_{obs}$  est dans la zone de rejet.

Sous  $H_0$ , l'échantillon tiré a une fréquence d'apparition inférieure à 5%. On refuse donc  $H_0$  et on choisit  $H_1$ .

## Remarques générales

**Test** : mécanisme permettant de trancher entre deux hypothèses à partir des résultats observés sur un ou plusieurs échantillons.

### Hypothèses

*Hypothèse nulle* : elle joue un rôle particulier ; elle affirme que les différences observées sont dues au hasard.

*Hypothèse alternative* : elle affirme que les différences sont significatives (en un sens à préciser).

### Les risques d'erreur

		Hypothèse vraie	
		$H_0$	$H_1$
Hypothèse retenue	$H_0$	$1 - \alpha$	$\beta$
	$H_1$	$\alpha$	$1 - \beta$

- $\alpha$  : seuil de significativité. C'est aussi la probabilité de rejeter  $H_0$  alors que  $H_0$  est vraie (risque de première espèce ou risque de commettre une erreur de type I)
- $\beta$  : risque de seconde espèce. C'est la probabilité d'accepter  $H_0$  alors que  $H_0$  est fautive (risque de commettre une erreur de type II).

$1 - \beta$  : probabilité de détecter correctement un cas où  $H_0$  doit être rejetée. Puissance du test.

**Illustrations** Commettre une ...

**Erreur de type I :** c'est voir une différence entre deux groupes alors qu'en fait, il n'y en a pas.

Exemples :

- Affirmer qu'un programme d'apprentissage coûteux a un effet sur le comportement des sujets, alors que c'est inexact
- "Mettre en évidence" une différence imaginaire entre les sexes, ou les races...

*Comment diminuer ce risque :* prendre  $\alpha$  petit, veiller à neutraliser les autres variables, etc

**Erreur de type II :** c'est ne pas voir de différence, alors qu'il y en a réellement une.

C'est souvent un moindre mal, mais...

Exemples :

- Ne pas mettre en évidence un risque de somnolence lié à l'absorption d'un médicament.
- L'usine de La Hague est-elle réellement inoffensive pour les riverains ?

*Comment diminuer ce risque :* augmenter la taille de l'échantillon, ne pas prendre  $\alpha$  trop petit, veiller à neutraliser les autres variables, bien choisir le test...

## Tests de comparaison de moyennes

### Comparaison de deux moyennes. Groupes appariés

On introduit le protocole dérivé des différences individuelles.

#### *Notations*

$\mu_1, \mu_2$  : moyennes respectives des deux variables étudiées

$\delta$  : moyenne des différences individuelles sur la population ( $\delta = \mu_1 - \mu_2$ ) (distribution normale)

$n$  : taille de l'échantillon

$\bar{x}_1, \bar{x}_2$  : moyennes respectives des deux variables sur un échantillon de taille  $n$

$\bar{d}$  : moyenne des différences individuelles sur un échantillon de taille  $n$  ( $\bar{d} = \bar{x}_1 - \bar{x}_2$ )

$s_c$  : écart type corrigé estimant l'écart type des différences individuelles sur la population parente

#### *Hypothèses du test*

$H_0$  :  $\mu_1 = \mu_2$ , c'est-à-dire  $\delta = 0$

$H_1$  : A choisir parmi :  $\mu_1 \neq \mu_2$  ou  $\mu_1 < \mu_2$  ou  $\mu_1 > \mu_2$

*Statistique de test.* Cas où  $n > 30$

$$Z = \frac{\bar{d}}{E} \text{ avec } E^2 = \frac{s_c^2}{n}$$

Z suit la loi normale centrée réduite.

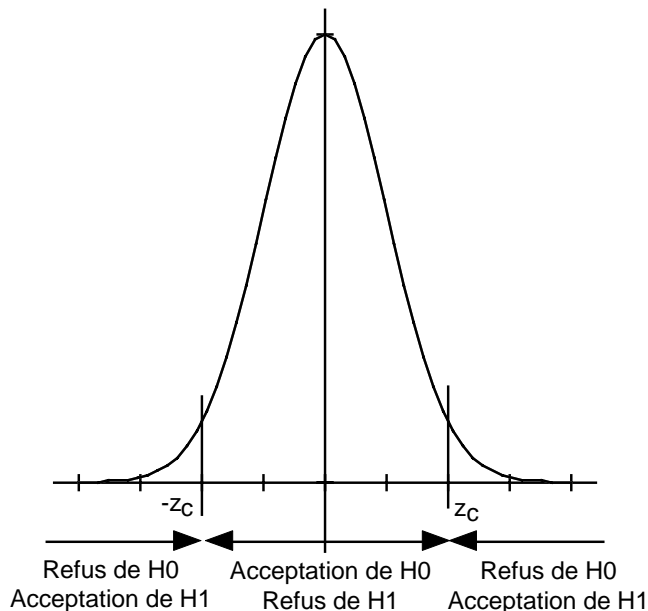
*Statistique de test.* Cas où  $n \leq 30$

$$T = \frac{\bar{d}}{E} \text{ avec } E^2 = \frac{s_c^2}{n}$$

T suit la loi de Student à  $n - 1$  ddl.

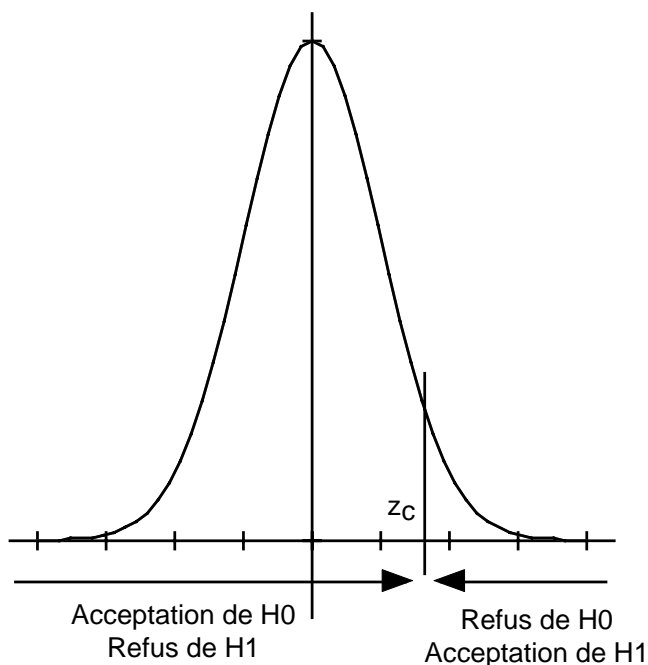
## Détermination de la règle de décision selon la forme de $H_1$

- Test bilatéral.  $H_1 : \mu_1 \neq \mu_2$



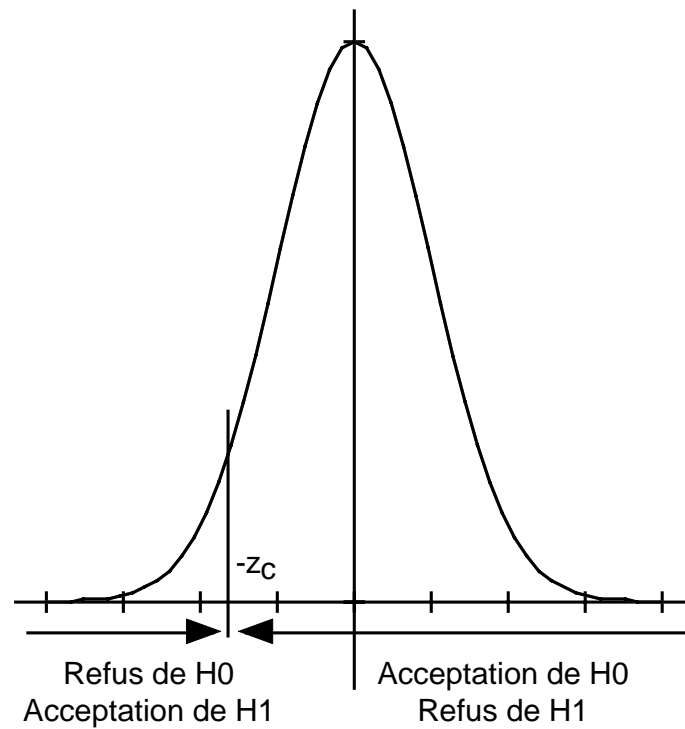
$$z_c \text{ définie par } P(-z_c \leq Z \leq z_c) = 1 - \alpha$$

- Test unilatéral.  $H_1 : \text{différence} > 0$



$$z_c \text{ définie par } P(Z \leq z_c) = 1 - \alpha$$

- Test unilatéral.  $H_1$  : différence  $< 0$

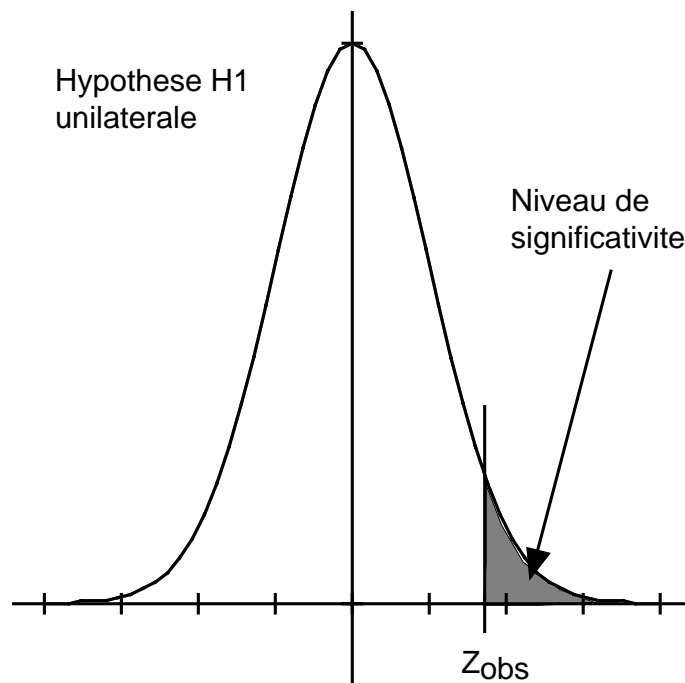


$z_c$  définie par  $P(Z \leq z_c) = 1 - \alpha$



## Remarque : raisonner en termes de niveau de significativité

Exemple pour une hypothèse  $H_1$  unilatérale “à droite” :



Soit  $\alpha$  le seuil et NivSig le niveau de significativité.

- Si  $NivSig < \alpha$ , on accepte  $H_1$ , et on refuse  $H_0$  au seuil choisi
- Si  $NivSig \geq \alpha$ , on refuse  $H_1$ , et on accepte  $H_0$  au seuil choisi

## **Comparaison de deux moyennes. Groupes indépendants**

### *Notations*

$\mu_1, \mu_2$  : moyennes sur les populations parentes respectives (distributions normales de même variance)

$n_1, n_2$  : tailles respectives des échantillons

$\bar{x}_1, \bar{x}_2$  : moyennes respectives sur des échantillons de tailles  $n_1$  et  $n_2$

$s_1, s_2$  : écarts types des deux échantillons.

$s_{1c}, s_{2c}$  : écarts types corrigés estimés à partir des échantillons.

### *Hypothèses du test*

$H_0 : \mu_1 = \mu_2$

$H_1$  : A choisir parmi :  $\mu_1 \neq \mu_2$  ou  $\mu_1 < \mu_2$  ou  $\mu_1 > \mu_2$

*Statistique de test.* Cas où  $n_1 > 30$  et  $n_2 > 30$

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{E} \text{ avec } E^2 = \frac{s_{1c}^2}{n_1} + \frac{s_{2c}^2}{n_2}$$

Sous  $H_0$ , Z suit la loi normale centrée réduite.

*Statistique de test.* Cas où  $n_1 \leq 30$  ou  $n_2 \leq 30$

$$T = \frac{\bar{x}_1 - \bar{x}_2}{E} \text{ avec } E^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2} \left( \frac{1}{n_1} + \frac{1}{n_2} \right)$$

T suit la loi de Student à  $n_1 + n_2 - 2$  ddl.

Si  $n_1 = n_2 (= n)$ , la formule se simplifie en :

$$T = \frac{\bar{x}_1 - \bar{x}_2}{E} \text{ avec } E^2 = \frac{s_{1c}^2 + s_{2c}^2}{n}$$

T suit la loi de Student à  $2(n - 1)$  ddl.

## Comparaison de deux proportions. Groupes indépendants

### *Notations*

$p_1, p_2$  : proportions dans les populations parentes respectives

$n_1, n_2$  : tailles respectives des échantillons

$f_1, f_2$  : proportions respectives dans des échantillons de tailles  $n_1$  et  $n_2$

### *Hypothèses du test*

$H_0 : p_1 = p_2$

$H_1$  : choisir entre :  $p_1 \neq p_2$  ou  $p_1 < p_2$  ou  $p_1 > p_2$

### *Statistique de test*

$$Z = \frac{f_1 - f_2}{E} \text{ avec}$$

$$E^2 = p(1 - p) \left( \frac{1}{n_1} + \frac{1}{n_2} \right) \text{ et } p = \frac{n_1 f_1 + n_2 f_2}{n_1 + n_2}$$

Si  $n_1 > 30$ ,  $n_2 > 30$  et  $p$  "ni trop grand, ni trop petit",  $Z$  suit la loi normale centrée réduite.

## Exemple

Variable sexe: deux groupes indépendants

Variable dépendante observée: succès/échec à une épreuve

	M	F	Ensemble
<b>Résultats</b>			
Succès	150	120	270
Echec	90	40	130
Total	240	160	400

## Calculs

$$f_1 = 62,5\%, f_2 = 75\%$$

$$p = 67,5\%, n_1 = 240, n_2 = 160$$

$$z_{obs} = -2.615$$

Pour un test bilatéral,  $z_{crit} = 1.96$

## Comparaison avec un test du $\chi^2$

Obs	Théo	Contrib
150	162	0,89
90	78	1.84
120	108	1.33
40	52	2.77
		6.84

$ddl = 1$ . Au seuil de 5%,  $\chi_{crit}^2 = 3.84$ .

Remarque:  $1.96^2 = 3.84$ ,  $(-2.615)^2 = 6.84$ , et ce n'est pas un hasard!

## Conclusion

### Tests *paramétriques*

#### Hypothèses *a priori* sur les populations parentes

- Groupes indépendants: distributions normales de même variance pour la variable dépendante,
- Groupes appariés: distribution normale des effets individuels dans la population parente ou échantillon de grande taille
- Comparaison de fréquences: échantillons de taille suffisante et fréquence “ni trop grande, ni trop petite”.

Il est généralement difficile de prouver que ces conditions sont respectées. Mais ces méthodes sont *robustes* et fournissent des résultats corrects même si les hypothèses ne sont qu’approximativement respectées.

Que fait-on quand elles ne le sont visiblement pas ?

## Tests non paramétriques

### Quel intérêt présentent ces tests ?

- En général, pas d'hypothèse a priori sur les populations parentes
- Ils peuvent s'appliquer à des variables ordinales (tests sur les rangs) ou même qualitatives (khi-2)

Il existe de nombreux tests non paramétriques. Nous n'étudierons que les plus courants.

## Test de la médiane sur des groupes indépendants

Une variable (la variable indépendante) définit deux groupes indépendants.

Une deuxième variable ordinale ou numérique.

### *Hypothèses*

$H_0$  : Les deux populations parentes ont même médiane.

$H_1$  : Les deux populations parentes ont des médianes différentes

### *Construction de la statistique de test*

On détermine la médiane  $M$  de la série obtenue en réunissant les deux échantillons.

On constitue un tableau de contingence en croisant la variable indépendante et la variable dérivée "position par rapport à  $M$ "

	Gr 1	Gr 2	Ensemble
$\leq M$	$N_{1\ 1}$	$N_{1\ 2}$	$N_{1.}$
$> M$	$N_{2\ 1}$	$N_{2\ 2}$	$N_{2.}$
Total	$N_{.1}$	$N_{.2}$	$N_{..}$

On fait un test du  $\chi^2$  sur le tableau obtenu.