Chapitre 9

Tests d'hypothèse sur deux moyennes

Tests sur deux moyennes

Comparer deux groupes:

Hommes vs femmes

Placebo vs médicament

Groupe contrôle vs groupe avec thérapie

Etc…

Deux moyennes : notations

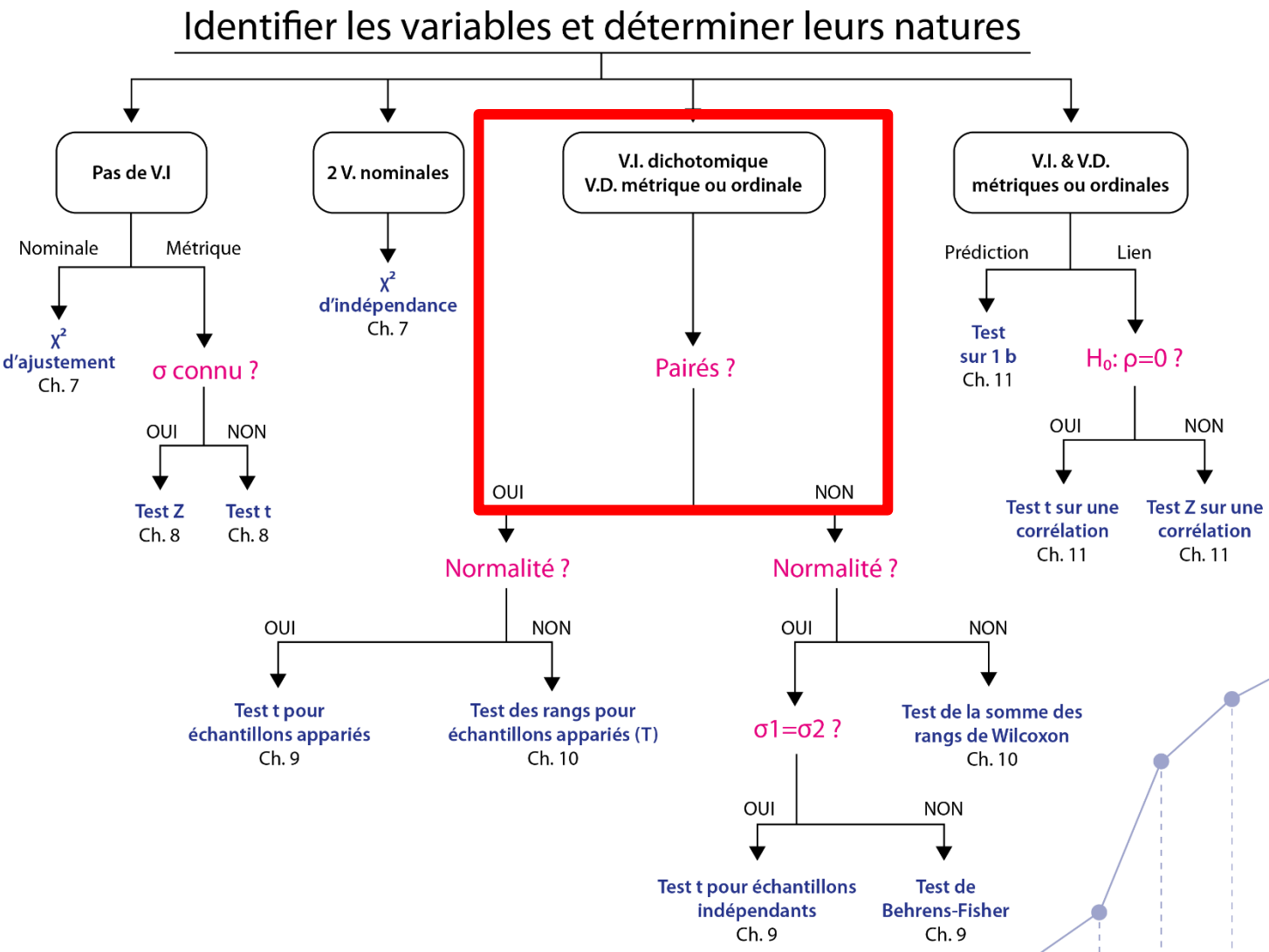
X_1	X_2
$\overline{X}_1$	$\overline{X}_2$
s_1	s_2
μ_1	μ_2
σ_1	σ_2

Echantillons pairés / non pairés

A distinguer:

- Echantillons pairés
- Echantillons non pairés

Choisir le bon test d'hypothèse



Echantillons pairés

Chaque score du premier échantillon est lié à un score du second échantillon

Souvent mesures répétées sur mêmes sujets

Echantillons pairés

	Avant	Après
Sujet 1	100	120
Sujet 2	35	45
Sujet 3	141	139
Sujet 4	45	56
Sujet 5	18	25
….	…	…

Echantillons pairés

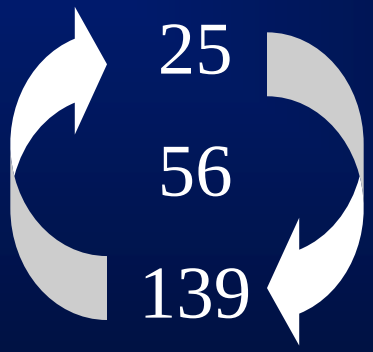
	Mari	Epouse
Couple 1	100	120
Couple 2	35	45
Couple 3	141	139
Couple 4	45	56
Couple 5	18	25
….	…	…

Echantillons non pairés

Hommes	Femmes
100	120
35	45
141	139
45	56
18	25
…	…

Echantillons non pairés

Hommes	Femmes
100	120
35	45
141	25
45	56
18	139
…	…



Exemple

Etude de Nurcombe et al. (1984):

Enfants PRN : mesure du développement mental à 6 mois et à 24 mois

Y a-t-il un changement significatif ?

Hypothèse nulle

$$H_0 : \mu_1 = \mu_2$$

$$H_A : \mu_1 \neq \mu_2$$

$$H_0 : \mu_1 - \mu_2 = 0$$

$$H_A : \mu_1 - \mu_2 \neq 0$$

Calcul des différences

	6 mois	24 mois	Différence
	124	114	-10
	94	88	-6
	115	102	-13
	110	127	17
	…	…	…
Moyenne	111,0	106,71	-4,29
Ecart-type	13,85	12,95	16,04
N	31	31	31

Scores D

Calcul des scores de différence (D)

Score + = amélioration

Score - = détérioration

Hypothèse nulle

Si les deux échantillons sont semblables: moyenne des D = zéro

Si $\mu_D = \mu_1 - \mu_2$ alors:

$H_0 : \mu_D = 0$

Formule

On se retrouve dans un test sur une moyenne avec H_0: μ = 0

$$t=\frac{\overline{X}-\mu}{\frac{S}{\sqrt{N}}}$$

devient

$$t=\frac{\overline{D}-\mu_D}{\frac{S_D}{\sqrt{N}}}$$

Exemple

	6 mois	24 mois	Différence
	124	114	-10
	94	88	-6
	115	102	-13
	110	127	17
	…	…	…
Moyenne	111,0	106,71	-4,29
Ecart-type	13,85	12,95	16,04
N	31	31	31

Exemple

$H_0 : \mu_D = 0$

$H_A : \mu_D \neq 0$

$\overline{D} = -4{,}29$ $S_D = 16{,}04$

$N = 31$ $dl = 31 - 1 = 30$

$\alpha = 0{,}05$ (par défaut)

Exemple

$$t_{obs} = \frac{\overline{D} - \mu_D}{\frac{S_D}{\sqrt{N}}} = \frac{-4{,}29 - 0}{\frac{16{,}04}{\sqrt{31}}} = \frac{-4{,}29}{2{,}88} = -1{,}49$$

$t_{0,025}$ = 2,042 (test bilatéral et 30 dl)

Exemple

1,49 < 2,042, ne pas rejeter H_0

NB: p = 0,0733 > 0,025

Nous ne pouvons pas affirmer que les scores de développement mental des enfants PRN évoluent entre 6 et 24 mois

Technique de l'intervalle de confiance

Intervalle de confiance pour μ_D

$$IC_{0,95} = \overline{X} \pm t_{\alpha/2} \frac{S}{\sqrt{N}}$$

devient

$$IC_{0,95} = \overline{D} \pm t_{\alpha/2} \frac{S_D}{\sqrt{N}}$$

Technique de l'intervalle de confiance

Intervalle de confiance pour notre exemple des enfants PRN

$$IC_{0,95} = -4,29 \pm 2,042 \frac{16,04}{\sqrt{31}} = -4,29 \pm 5,88$$

$$-10,17 < \mu_D < 1,59$$

Technique de l'intervalle de confiance

$$-10{,}17 < \mu_D < 1{,}59$$

Puisque zéro se trouve à l'intérieur de l'intervalle de confiance, nous ne pouvons exclure que le score de développement mental n'évolue pas entre 6 et 24 mois.

Tester une différence autre que zéro

Hypothèse: les enfants PRN présentent une détérioration de leur développement mental de plus de 15 points

$H_0 : \mu_D = -15$

$H_A : \mu_D < -15$

Tester une différence autre que zéro

$H_0 : \mu_D = -15$

$H_A : \mu_D < -15$

$$t_{obs} = \frac{\overline{D} - \mu_D}{\frac{S_D}{\sqrt{N}}} = \frac{-4{,}29 - (-15)}{\frac{16{,}04}{\sqrt{31}}} = \frac{10{,}71}{2{,}88} = 3{,}72$$

Tests unilatéraux et bilatéraux

Quand faut-il rejeter H_0:

Test t unilatéral $\mu_D >$: $t_{obs} > t_{0,05}$

Test t unilatéral $\mu_D <$: $t_{obs} < -t_{0,05}$

Test t bilatéral : valeur absolue de $t_{obs} > t_{0,025}$

Données manquantes

1. Pourquoi données manquantes ?

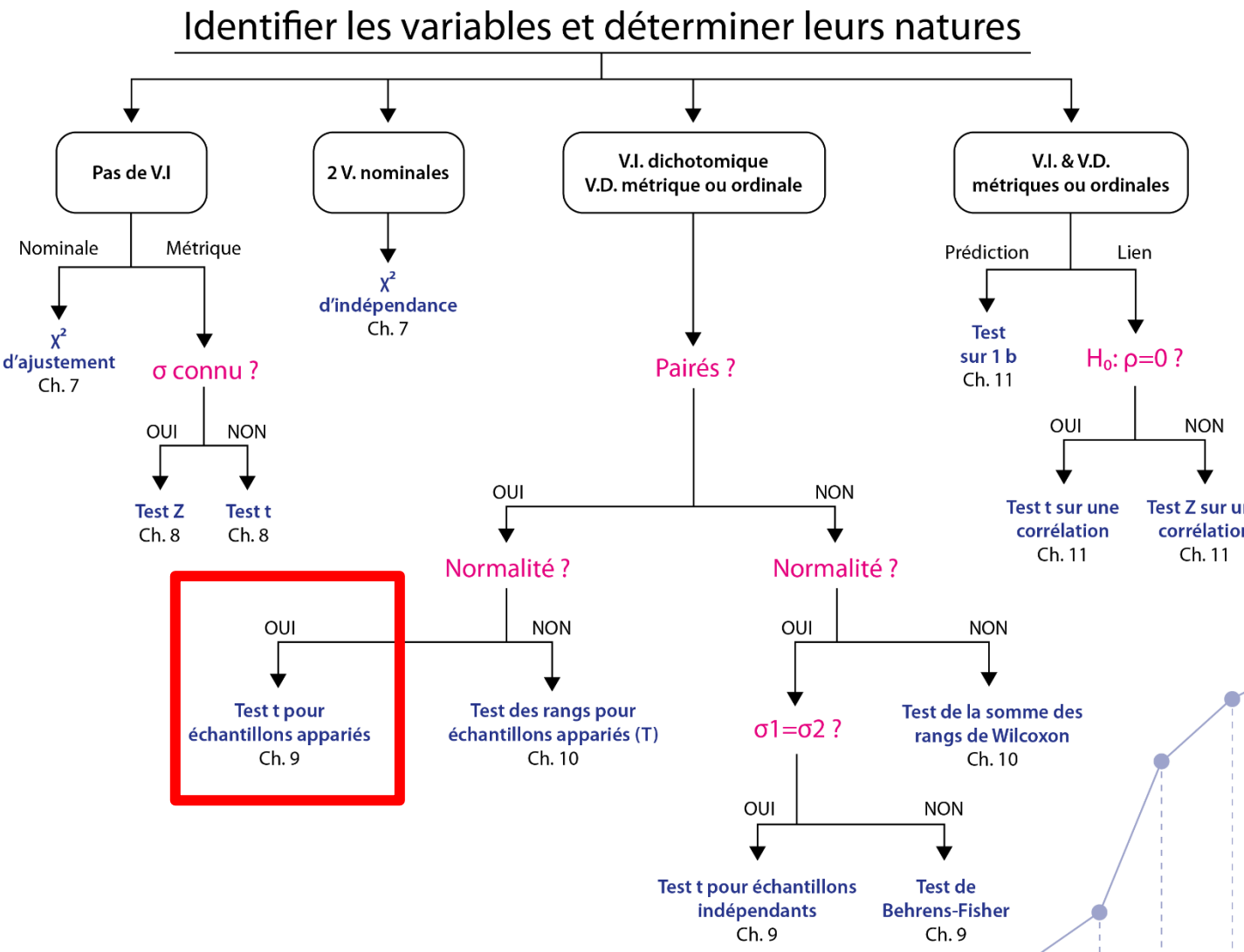
2. Eliminer la paire

3. Tests alternatifs possibles

Conditions d'application du test t sur échantillons pairés

Les scores de différence proviennent d'une population normalement distribuée

Choisir le bon test d'hypothèse



Tests sur deux moyennes pour échantillons non pairés

Echantillons non pairés

Hommes	Femmes	Différences
100	120	20
35	45	10
141	139	-2
45	56	11
18	25	7
	Moyenne	9,2
	SD	7,92

Echantillons non pairés

Hommes	Femmes	Différences
100	56	-44
35	25	-10
141	120	-21
45	45	0
18	139	121
	Moyenne	9,2
	SD	64,6

Echantillons non pairés

A distinguer:

- Variances homogènes
  - Test t classique
- Variances hétérogènes
  - Test de Behrens-Fisher

Exemple avec variances homogènes

Etude sur la mémoire de sujets jeunes ou âgés (Eysenck, 1974)

Liste de mots à mémoriser

VI: Sujets jeunes ou âgés

VD: Nombre de mots rappelés

Exemple avec variances homogènes

Sujets jeunes					Sujets âgés				
21	19	17	15	22	10	19	14	5	10
16	22	22	18	21	11	14	15	11	11

$\overline{X}_1 = 19,3$	$\overline{X}_2 = 12,0$
$S_1^2 = 7,122$	$S_2^2 = 14,000$
$N_1 = 10$	$N_2 = 10$

Exemple avec variances homogènes

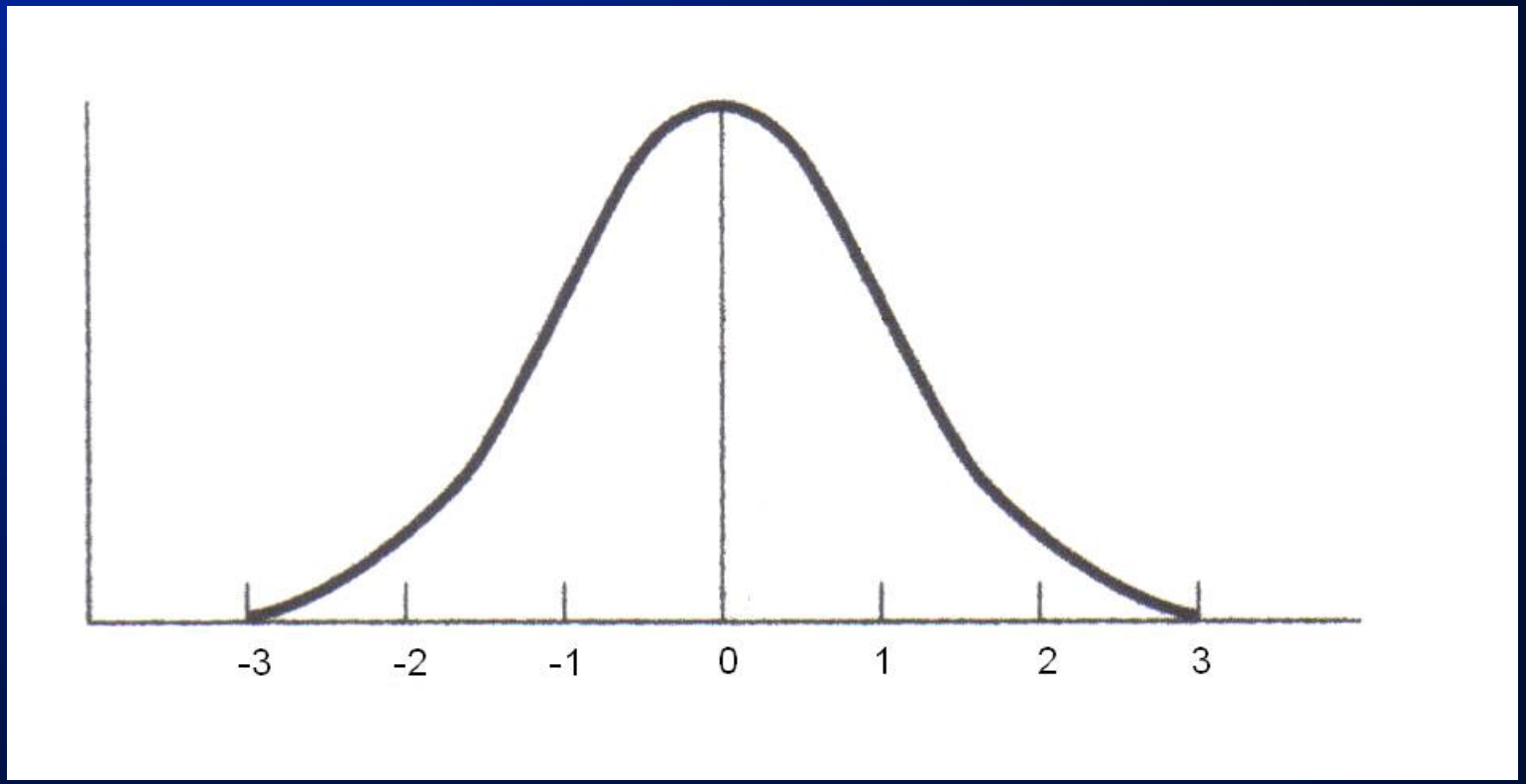
$H_0 : \mu_1 - \mu_2 = 0$

$H_A : \mu_1 - \mu_2 \neq 0$

Nous avons obtenu $\overline{X}_1 - \overline{X}_2 = 7,3$

Erreur d'échantillonnage ou véritable différence ?

La distribution d'échantillonnage de différences entre les moyennes



7,3 peut-il appartenir à cette distribution ?

La distribution d'échantillonnage de différences entre les moyennes

Pour calculer les probabilités, il faut connaître les paramètres de cette distribution

La distribution d'échantillonnage de différences entre les moyennes

Moyenne = $\mu_1 - \mu_2$

<u>Loi de la somme des variances</u> :

« la variance d'une somme ou d'une différence de deux variables indépendantes est égale à la somme de leurs variances »

La distribution d'échantillonnage de différences entre les moyennes

Moyenne = $\mu_1 - \mu_2$

La variance d'une distribution d'échantillonnages de moyennes est égale à σ^2/N

σ_1^2/N_1 **pour la population X_1**

σ_2^2/N_2 **pour la population X_2**

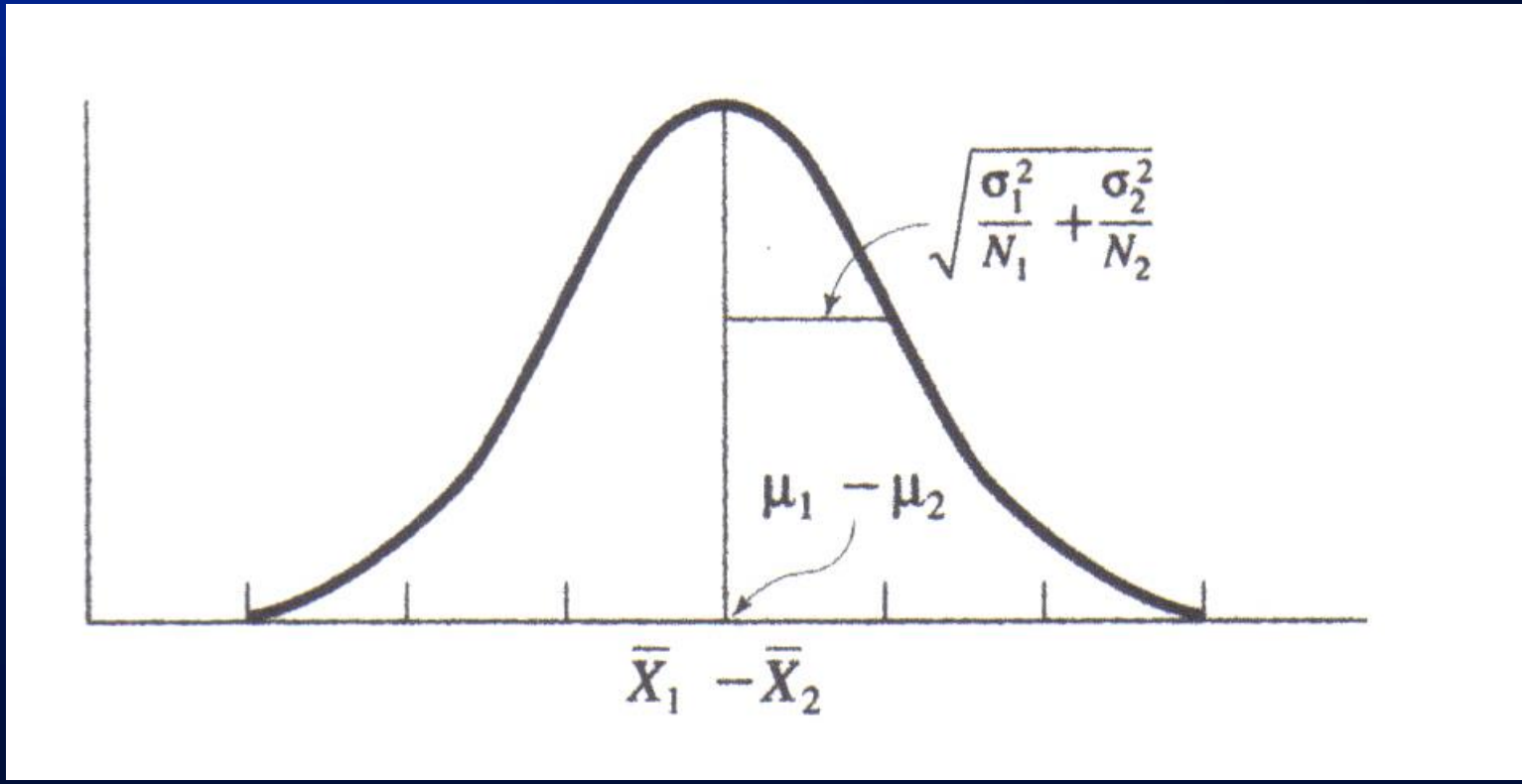
La distribution d'échantillonnage de différences entre les moyennes

Moyenne = $\mu_1 - \mu_2$

$$\sigma^2_{\bar{X}_1 - \bar{X}_2} = \frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}$$

La somme ou la différence de deux variables indépendantes normalement distribuées est elle aussi normalement distribuée

La distribution d'échantillonnage de différences entre les moyennes



La distribution d'échantillonnage de différences entre les moyennes

Dernier problème : Les variances des populations sont généralement inconnues

Il faut les estimer par S_1^2 **et** S_2^2

La distribution d'échantillonnage de différences entre les moyennes

$$t = \frac{\left(\overline{X}_1 - \overline{X}_2\right) - \left(\mu_1 - \mu_2\right)}{\sqrt{\frac{S_1^2}{N_1} + \frac{S_2^2}{N_2}}}$$

La distribution d'échantillonnage de différences entre les moyennes

La formule peut encore être améliorée

Nous testons deux groupes avec variances homogènes.

Donc: $\sigma_1^2 = \sigma_2^2 = \sigma^2$

La distribution d'échantillonnage de différences entre les moyennes

Puisque $\sigma_1^2 = \sigma_2^2 = \sigma^2$

La moyenne de S_1^2 **et** S_2^2

serait une meilleure estimation de σ^2

Estimation combinée de la variance

Calcul d'une moyenne pondérée de la variance selon les dl

$$S_P^2 = \frac{(N_1 - 1)}{N_1 + N_2 - 2} S_1^2 + \frac{(N_2 - 1)}{N_1 + N_2 - 2} S_2^2$$

$$S_P^2 = \frac{(N_1 - 1)S_1^2 + (N_2 - 1)S_2^2}{N_1 + N_2 - 2}$$

Test t sur deux échantillons indépendants

$$t=\frac{\left(\overline{X}_1-\overline{X}_2\right)-\left(\mu_1-\mu_2\right)}{\sqrt{\frac{S_P^2}{N_1}+\frac{S_P^2}{N_2}}}=\frac{\left(\overline{X}_1-\overline{X}_2\right)-\left(\mu_1-\mu_2\right)}{\sqrt{S_P^2\left(\frac{1}{N_1}+\frac{1}{N_2}\right)}}$$

Les degrés de liberté

$$t = \frac{\left(\overline{X}_1 - \overline{X}_2\right) - \left(\mu_1 - \mu_2\right)}{\sqrt{S_P^2\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}}$$

$dl = N_1 + N_2 - 2$

Résolution du test

Sujets jeunes					Sujets âgés				
21	19	17	15	22	10	19	14	5	10
16	22	22	18	21	11	14	15	11	11

$\overline{X}_1 = 19,3$	$\overline{X}_2 = 12,0$
$S_1^2 = 7,122$	$S_2^2 = 14,000$
$N_1 = 10$	$N_2 = 10$

Résolution du test

$H_0 : \mu_1 - \mu_2 = 0$

$H_A : \mu_1 - \mu_2 \neq 0$

dl = 10 + 10 –2 = 18

α = 0,05 (par défaut)

Résolution du test

$$S_P^2 = \frac{(N_1 - 1)S_1^2 + (N_2 - 1)S_2^2}{N_1 + N_2 - 2} = \frac{(9 \times 7{,}122) + (9 \times 14)}{18} = 10{,}561$$

$$t_{obs} = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{S_P^2\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}} = \frac{19{,}3 - 12}{\sqrt{10{,}561\left(\frac{1}{10} + \frac{1}{10}\right)}} = \frac{7{,}3}{1{,}4533} = 5{,}02$$

Résolution du test

$t_{0,025} = 2,101$ (test bilatéral et 18 dl)

Puisque $5,02 > 2,101$, rejeter H_0

Les sujets âgés se rappellent en moyenne moins de mots que les sujets jeunes.

Tests unilatéraux et bilatéraux

Quand faut-il rejeter H_0:

Test t unilatéral > : $t_{obs} > t_{0,05}$

Test t unilatéral < : $t_{obs} < -t_{0,05}$

Test t bilatéral : valeur absolue de $t_{obs} > t_{0,025}$

Technique de l'intervalle de confiance

Intervalle de confiance pour $\mu_1 - \mu_2$

$$IC_{0,95} = \overline{X} \pm t_{\alpha/2} \frac{S}{\sqrt{N}}$$

devient

$$IC_{0,95} = \left(\overline{X}_1 - \overline{X}_2\right) \pm t_{\alpha/2} \sqrt{\frac{S_P^2}{N_1} + \frac{S_P^2}{N_2}}$$

Technique de l'intervalle de confiance

Intervalle de confiance pour l'exemple des sujets jeunes et âgés

$$\overline{X}_1 - \overline{X}_2 = 19,3 - 12 = 7,3$$

$$S_P^2 = 10,561$$

N_1 = 10 **N_2 = 10**

$t_{0,025}$ pour un test bilatéral et 18 dl = 2,101

Technique de l'intervalle de confiance

$$IC_{0,95} = 7,3 \pm 2,101 \sqrt{\frac{10,561}{10} + \frac{10,561}{10}} = 7,3 \pm 3,05$$

$4,25 < \mu_1 - \mu_2 < 10,35$

Puisque zéro ne se trouve pas dans l'intervalle de confiance, nous pouvons conclure que les jeunes retiennent significativement plus de mots que les sujets âgés

Conditions d'application

1) Normalité de la distribution

   Sinon: test non paramétrique

2) Homogénéité des variances

   Test robuste

   Si inégalité des variances + $N_1 \neq N_2$:

   Solution de Behrens-Fisher

Le test de Behrens-Fisher

A utiliser si : **- Hétérogénéité des variances**

- $N_1 \neq N_2$

Puisque variances hétérogènes:

On ne peut plus calculer S_P^2

Le test de Behrens-Fisher

On revient à:

$$t' = \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{N_1} + \frac{S_2^2}{N_2}}}$$

Le test de Behrens-Fisher

t' n'est pas distribué comme le t avec $N_1 + N_2 - 2$ dl

$$dl' = \frac{\left(\frac{S_1^2}{N_1} + \frac{S_2^2}{N_2}\right)^2}{\frac{\left(\frac{S_1^2}{N_1}\right)^2}{N_1 - 1} + \frac{\left(\frac{S_2^2}{N_2}\right)^2}{N_2 - 1}}$$

Arrondir dl' à l'entier le plus proche

Exercice

Gross (1984) a étudié l'évolution du poids chez les personnes victimes de boulimie

Deux groupes: - Boulimie simple
- Boulimie avec vômissements

VD: écart, en pour cent, du poids par rapport à des sujets normaux

Exercice

Boulimie simple : $\overline{X}_1 = 4{,}61$

$S_1^2 = 219{,}04$

N_1 = 49

Boulimie avec vomissements : $\overline{X}_2 = -0{,}83$

$S_2^2 = 79{,}21$

N_2 = 32

Exercice

$H_0 : \mu_1 - \mu_2 = 0$

$H_A : \mu_1 - \mu_2 \neq 0$

α = 0,05 (par défaut)

$$t'_{obs} = \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{N_1} + \frac{S_2^2}{N_2}}} = \frac{4{,}61 - (-0{,}83)}{\sqrt{\frac{219{,}04}{49} + \frac{79{,}21}{32}}} = \frac{5{,}44}{\sqrt{6{,}9455}} = 2{,}064$$

Exercice

$$dl' = \frac{\left(\frac{S_1^2}{N_1} + \frac{S_2^2}{N_2}\right)^2}{\frac{\left(\frac{S_1^2}{N_1}\right)^2}{N_1 - 1} + \frac{\left(\frac{S_2^2}{N_2}\right)^2}{N_2 - 1}} = \frac{\left(\frac{219{,}04}{49} + \frac{79{,}21}{32}\right)^2}{\frac{\left(\frac{219{,}04}{49}\right)^2}{48} + \frac{\left(\frac{79{,}21}{32}\right)^2}{31}}$$

$$= \frac{48{,}2400}{0{,}4163 + 0{,}1976} = 78{,}58$$

Exercice

$t'_{obs} = 2{,}064$

$dl' = 79$

$t_{0{,}025} = ?$

Exercice

$t'_{obs} = 2,064$

$t_{0,025} = 2,009$ (arrondi à 50, la valeur inférieure la plus proche de 79)

Puisque $2,064 > 2,009$, rejeter H_0

Il existe une différence significative entre les poids des deux types de boulimie

Tester l'égalité des variances

Comment savoir si on est confronté à un problème d'hétérogénéité des variances puisque nous ne connaissons pas les variances des populations ?

Recours à un test statistique

Tester l'égalité des variances

Le test F d'égalité des variances

$$F = \frac{S_1^2}{S_2^2} \quad \textbf{ou} \quad F = \frac{S_2^2}{S_1^2}$$

Plus grande des deux variances au numérateur

Exercice

Test d'égalité des variances pour l'exemple sur les boulimies

$S_1^2 = 219{,}04$ **$N_1 = 49$**

$S_2^2 = 79{,}21$ **$N_2 = 32$**

Exercice

$H_0 : \quad \sigma_1^2 = \sigma_2^2$

$H_A : \quad \sigma_1^2 \neq \sigma_2^2$

α = 0,05 (par défaut)

Exercice

$$F_{obs} = \frac{S_1^2}{S_2^2} = \frac{219{,}04}{79{,}21} = 2{,}765$$

$dl_{numérateur} = 49 - 1 = 48$

$dl_{dénominateur} = 32 - 1 = 31$

	Degrés de liberté du numérateur							
	9	10	15	20	25	30	40	50
28	2.24	2.19	2.04	1.96	1.91	1.87	1.82	1.79
30	2.21	2.16	2.01	1.93	1.88	1.84	1.79	1.76
40	2.12	2.08	1.92	1.84	1.78	1.74	1.69	1.66
50	2.07	2.03	1.87	1.78	1.73	1.69	1.63	1.60

Exercice

$F_{0,05\ ;\ 50\ ;\ 30} = 1,76$

Puisque 2,765 > 1,76, rejeter H_0

Les deux échantillons proviennent de population dont les variances sont différentes

Choisir le bon test d'hypothèse