Lois de distributions

La loi normale

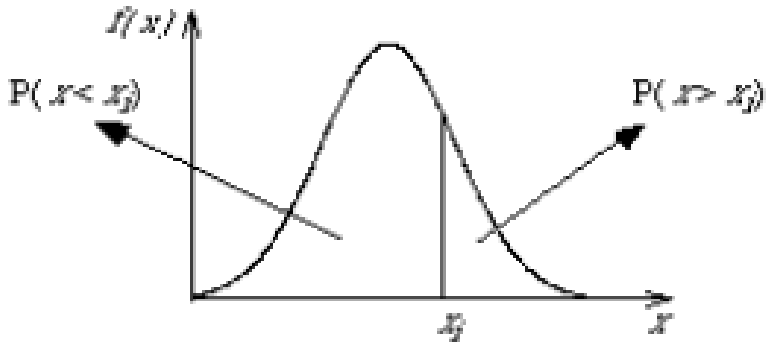
La loi normale repose sur l'estimation de deux paramètres de la population statistique:

- **la moyenne μ**
- **l'écart type σ**

La courbe (appelée "fonction de densité de probabilité") a la formule suivante:

$$f_{(x)} = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

La probabilité qu'une variable *x* prenne une valeur plus petite ou plus grande qu'une certaine valeur x_i s'obtient en calculant l'aire sous la courbe:



La loi normale centrée-réduite *N*(0,1)

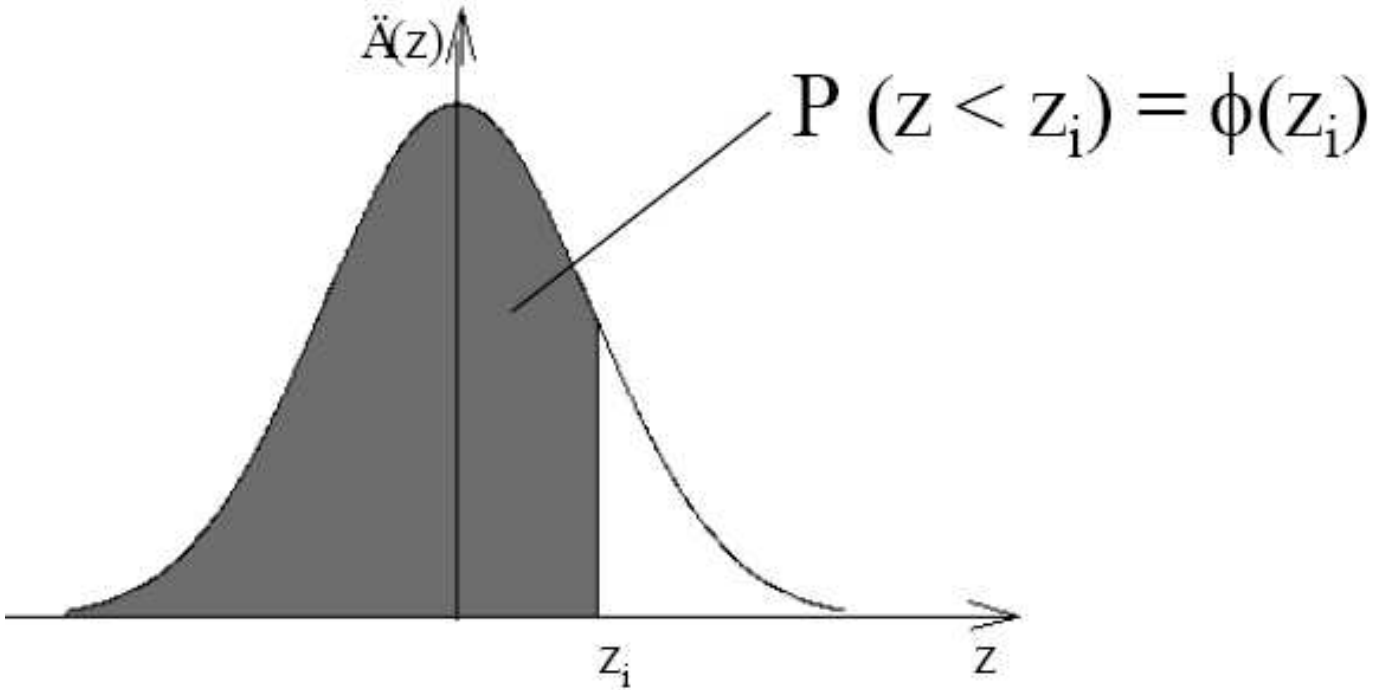
Il s'agit d'une loi **loi normale pour laquelle toutes les valeurs x_i sont centrées-réduites**:

$$z_i = \frac{x_i - \mu}{\sigma}$$

Propriétés

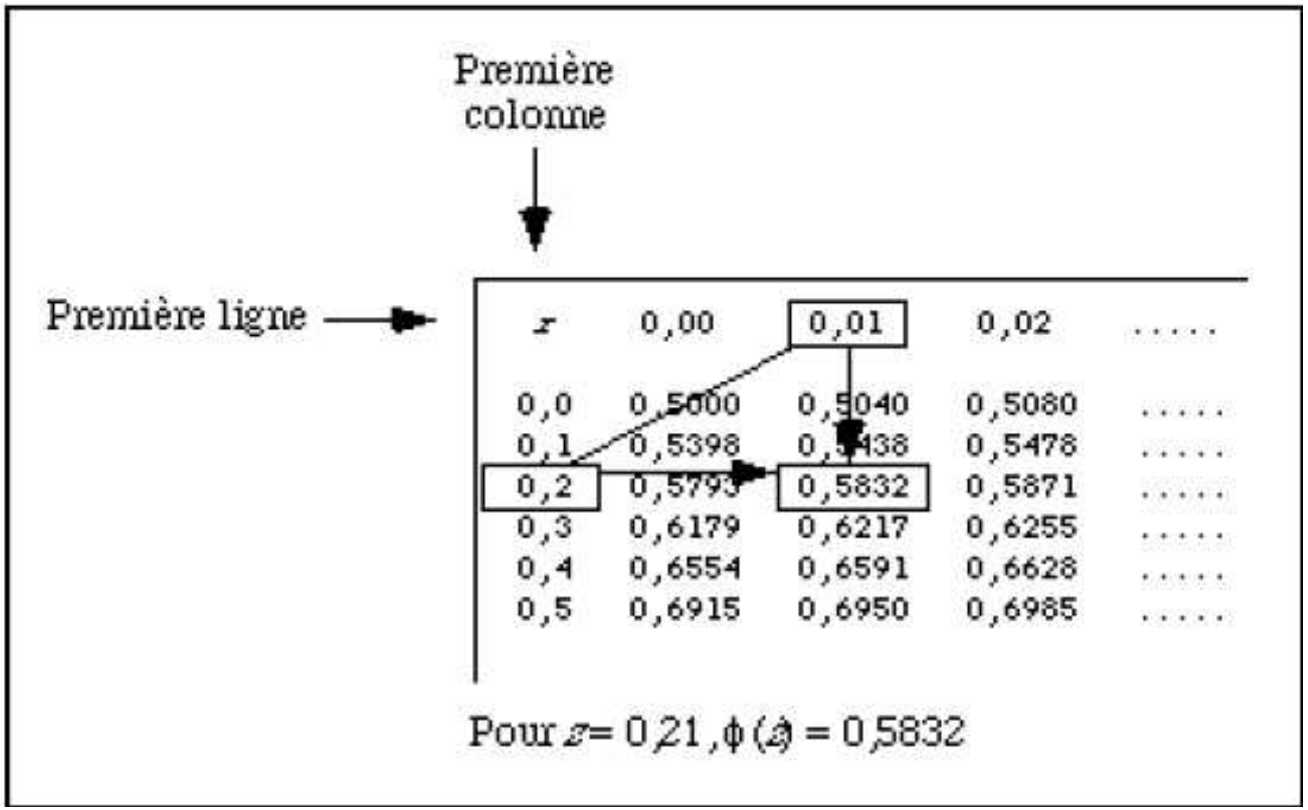
- μ = 0 et σ = 1
- Il n'y a pas d'unités
- L'aire totale sous la courbe = 1 (donc l'aire pour *z* allant de moins l'infini à zéro = 0.5)
- La courbe est parfaitement symétrique: $f(z) = f(-z)$ (donc l'aire pour *z* allant de moins l'infini à zéro = l'aire pour *z* allant de 0 à plus l'infini = 0.5)
- La courbe est continue, donc P(*z*) = 0
- Les probabilités correspondent directement à la surface sous la courbe

Usage le plus simple: on veut trouver la probabilité qu'une valeur *z* soit inférieure à une valeur limite z_i
Cela correspond à la surface de la courbe normale centrée réduite située à gauche du trait vertical marquant la position de la valeur limite z_i :

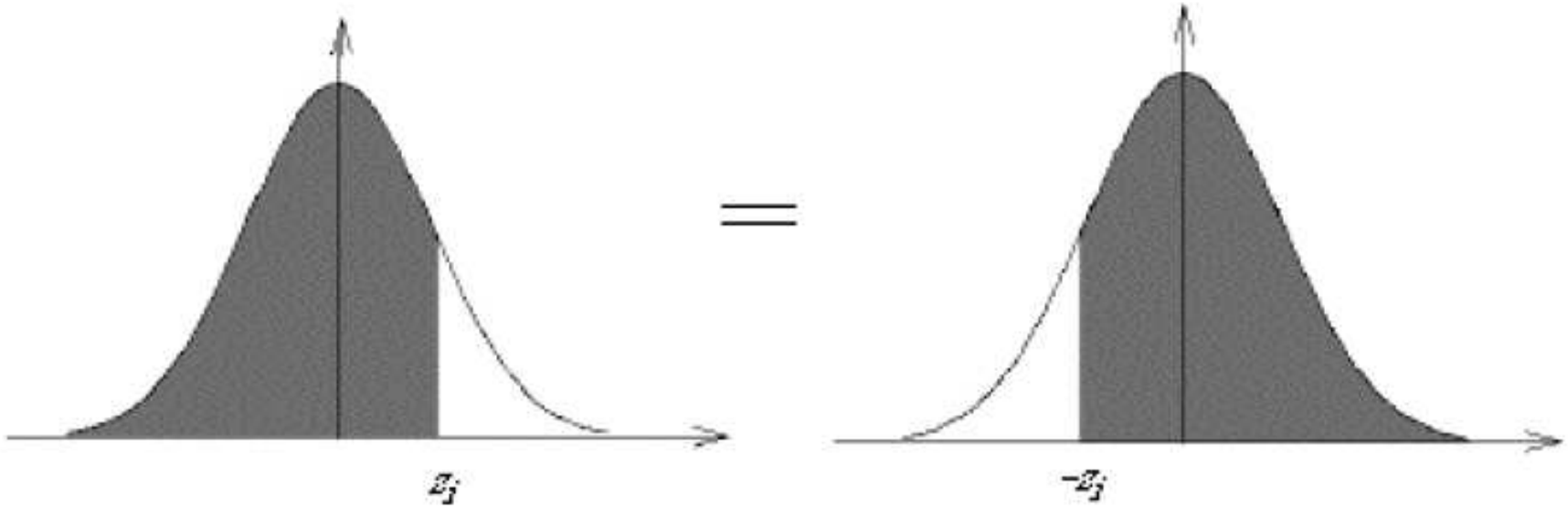


Exemple: probabilité que z soit inférieur à z_i = +0,21 [qui s'exprime P (z < 0,21)]:

La *première colonne* de la table indique *l'unité* et la *première décimale* de z_i
La *première ligne* de la table indique la *deuxième décimale* de z_i :



La table ne contient que les valeurs de z_i positives parce que:



P (z < 0,21) = 0,5832
P (z > –0,21) = 0,5832
P (0 < z < 0,21) = 0,5832 – 0,5000 = 0,0832
P (–0.21 < z < 0) = 0,5832 – 0,5000 = 0,0832
P (–0,21 < z < 0,21) = 0,0832 + 0,0832 = 0,1664

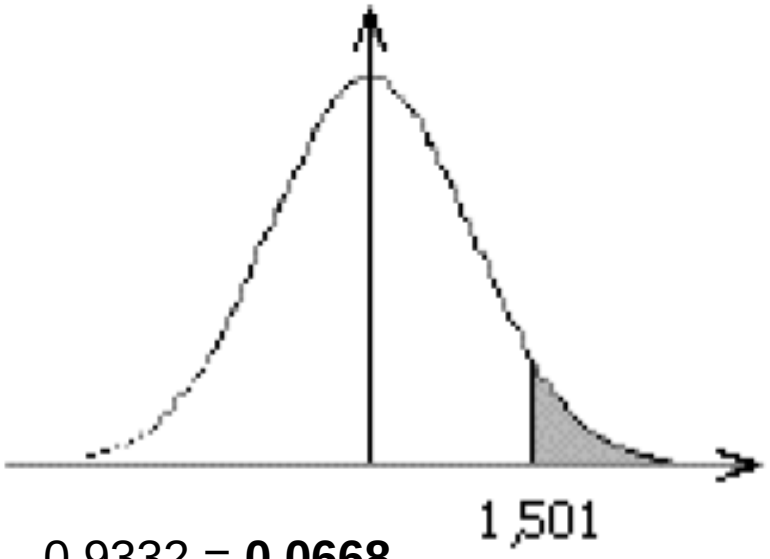
Robert va à la pêche

Le beau-frère de Robert a pêché un brochet de 538,9 mm dans le lac Raymond. En admettant que la longueur des brochets de ce lac suit une loi normale *N*(467 mm, 47,9 mm), quelle est la probabilité que Robert pêche un brochet **plus long** que celui de son beau-frère?

1) Traduction: P (x > 538,9mm) = ?

2) Transformons 538,9 mm en z: $$z = \frac{538,9\text{ mm} - 467,0\text{ mm}}{47,9\text{ mm}} = 1,501$$

3) Donc: P (x > 538,9 mm) = P (z > 1,501):



P(z > 1,501) = 1 – P(z < 1,501) = 1 – 0,9332 = **0,0668**

La probabilité que Robert pêche un brochet plus long que celui de son beau-frère est donc de 0,0668.

Robert retourne à la pêche!

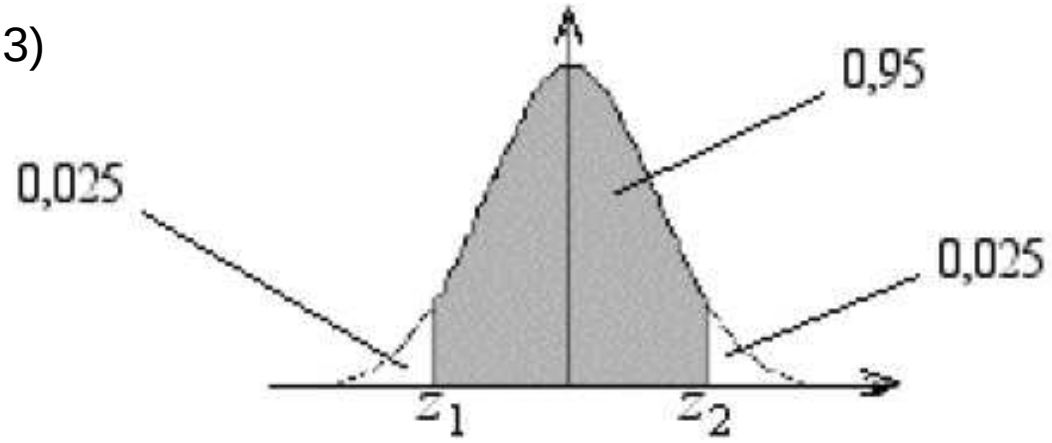
Si la longueur des brochets du lac Raymond suit une loi normale N(467 mm, 47,9 mm), entre quelles valeurs se situent 95 % des longueurs des brochets de ce lac ?

1) Traduction: P ($x_1 < x < x_2$) = 0,95

2) Transformons x_1 et x_2 en z: P ($z_1 < z < z_2$) = 0,95

Le problème est donc inverse du précédent: nous avons la probabilité mais pas z

3)



P ($z < z_2$) = 0,95 + 0,025 = 0,975

z_2 = 1,96 On cherche 0.975 dans le corps de la table (φ = 0.9750), et on remonte dans les marges pour connaître le *z* correspondant!
z_1 = –1,96 par symétrie

4) Retransformons z_1 et z_2 en *x* par l'opération inverse d'un centrage-réduction on multiplie *z* par l'écart type de la variable, puis on ajoute la moyenne au résultat :

$x_1 = (z_1 \times 47{,}9\text{ mm}) + 467\text{ mm} = (-1{,}96 \times 47{,}9\text{ mm}) + 467\text{ mm} = 373{,}12\text{ mm}$
$x_2 = (z_2 \times 47{,}9\text{ mm}) + 467\text{ mm} = (1{,}96 \times 47{,}9\text{ mm}) + 467\text{ mm} = 560{,}88\text{ mm}$

Ainsi, **95%** des brochets du lac Raymond ont une longueur comprise entre **373,12 mm et 560,88 mm.**

Robert veut un gros brochet !

Robert se fait dire que les brochets du lac Abitibi sont plus longs que ceux du lac Raymond :

	$\bar{x}$	s_x
Lac Raymond	467,0 mm	47,9 mm
Lac Abitibi	481,0 mm	44,8 mm

Il est très probable que certaines tailles de brochets sont représentées dans les deux lacs. Mais au fait...

A. Quel est le chevauchement entre les deux distributions ?

Considérons les intervalles comprenant 99% des brochets dans chacun des lacs, soit ceux compris entre $z_1 = -2{,}575$ et $z_2 = 2{,}575$ dans le cas de la loi $N(0,1)$:

Lac Raymond:

$x_1 = (-2{,}575 \times 47{,}9) + 467 = 343{,}7$ mm

$x_2 = (2{,}575 \times 47{,}9) + 467 = 590{,}3$ mm

Lac Abitibi:

$x_1 = (-2{,}575 \times 44{,}8) + 481 = 365{,}6$ mm

$x_2 = (2{,}575 \times 44{,}8) + 481 = 596{,}4$ mm

1) Traduction: P (x > 590,3 mm) = ?

2) Transformons 590,3 mm en z : z = (590,3 mm - 481 mm)/44.8 mm = 2,459

3) Donc: P (x > 590,3 mm) = P (z > 2,46)

4) P (z > 2,44) = 1 – P (z < 2,46) = 1 – 0,9931 = **0,0069**

Même si les plus gros brochets se trouvent effectivement dans le lac Abitibi, ils sont visiblement très rares (ils semblent constituer 0,69% des brochets de ce lac). Pas de quoi monter une expédition...

Calcul d’une distribution de probabilités obéissant à la loi normale

Jusqu'ici nous avons passé de la valeur réelle des données à la distribution normale pour quelques valeurs isolées. Nous allons voir ici comment construire une distribution normale dont les paramètres (moyenne et écart type) sont les mêmes que ceux de nos données. Il y a deux méthodes, celle des surfaces et celles des ordonnées.

1. Ce dont nous avons besoin

• La distribution de fréquences de la variable (qui a donc été divisée en classes)
• la moyenne x de l'échantillon ou μ de la population statistique
• l'écart type s_x de l'échantillon ou σ de la population statistique.

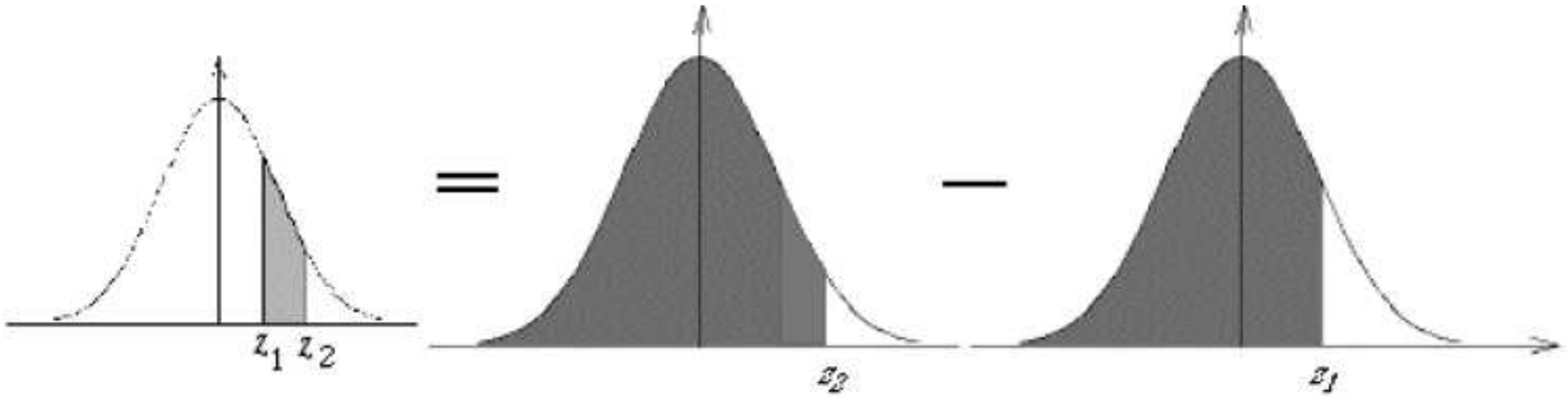
2. Principe

Trouver la probabilité associée à chacune des classes.

3. Méthodes

a) Méthode des surfaces

• On transforme les ***bornes*** de chaque classes en z (en centrant-réduisant les bornes) et on utilise la table III pour déterminer la probabilité associée à chacun de ces intervalles.



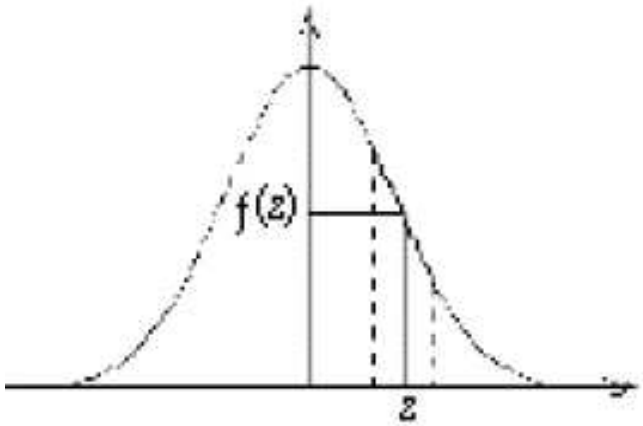
$$P\,(x_1 < x < x_2) = P\,(z_1 < z < z_2) = \left| P\,(z < z_2) - P\,(z < z_1) \right|$$

b) Méthode des ordonnées

• On transforme les ***indices*** de chaque classe en *z* et on utilise la table IV pour déterminer l'ordonnée (la fréquence relative) associée à chacun de ces indices selon la loi normale centré réduite.

• On utilise une formule pour déterminer la probabilité de chaque classe à partir de son intervalle et de son ordonnée:

$$P(x_1 < x < x_2) = \frac{(x_2 - x_1) \times f(z)}{s_x}$$



Cette formule fait donc intervenir :

-($x_2 – x_1$) l'intervalle de classe (en données brutes, non centrées réduites!)

-*f(z)* la fréquence relative correspondant à la valeur *z* (table IV)

- s_x l'écart type de l'échantillon.

Exemple : Robert va à la chasse aux gélinottes.

Distribution de fréquences absolues de la longueur (mm) de la rectrice centrale (x_i) de 592 gélinotes huppées mâles à un moment donné.

x_i	110	120	130	140	150	160	170	180	190	200
f	3	26	72	103	117	97	72	60	31	11

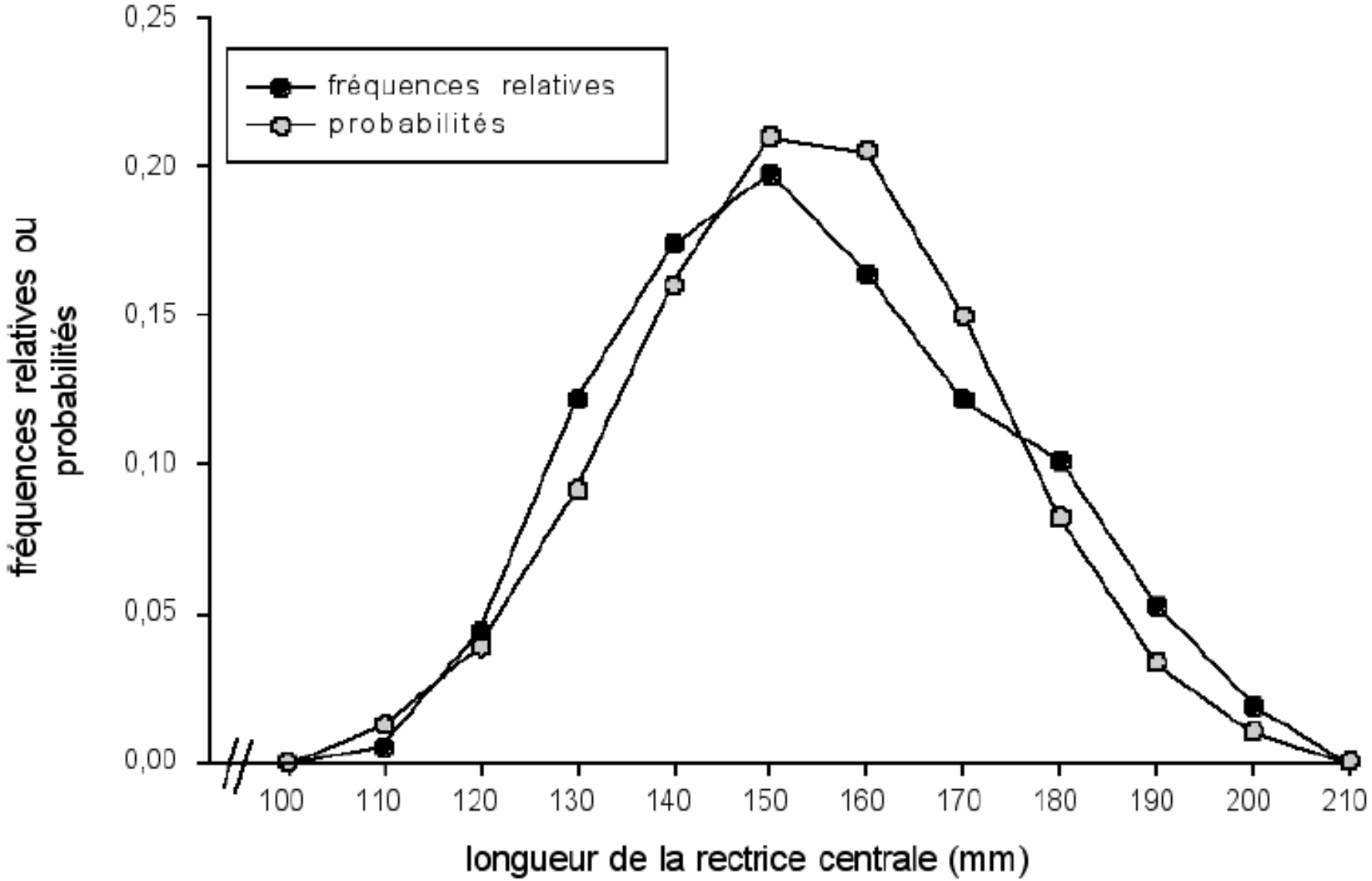
Sachant que ***moyenne* = 154,3 mm** et que $\boldsymbol{s_x}$ **= 18,5 mm**, quelle serait la distribution de probabilités de cette variable selon la loi normale ?

A) Méthode des surfaces

Indices	f	borne inf	z_{inf}	$P(z < z_{inf})$	borne su	z_{sup}	$P(z < z_{sup})$	$P(z_{inf} < z < z_{sup})$
110	3	105	-2,66	0,0039	115	-2,12	0,0170	0,0131
120	26	115	-2,12	0,0170	125	-1,58	0,0571	0,0401
130	72	125	-1,58	0,0571	135	-1,04	0,1492	0,0921
140	103	135	-1,04	0,1492	145	-0,50	0,3085	0,1593
150	117	145	-0,50	0,3085	155	0,04	0,5160	0,2075
160	97	155	0,04	0,5160	165	0,58	0,7190	0,2030
170	72	165	0,58	0,7190	175	1,12	0,8686	0,1496
180	60	175	1,12	0,8686	185	1,66	0,9515	0,0829
190	31	185	1,66	0,9515	195	2,20	0,9861	0,0346
200	11	195	2,20	0,9861	205	2,74	0,9969	0,0108
somme	592							0,9930

B. Méthode des ordonnées

Indices	f	z	f(z)	$x_2 - x_1$	$P(x_1 < x < x_2)$
110	3	-2,39	0,0229	10	0,0124
120	26	-1,85	0,0721	10	0,0390
130	72	-1,31	0,1691	10	0,0914
140	103	-0,77	0,2966	10	0,1603
150	117	-0,23	0,3885	10	0,2100
160	97	0,31	0,3802	10	0,2055
170	72	0,85	0,2780	10	0,1503
180	60	1,39	0,1518	10	0,0821
190	31	1,93	0,0620	10	0,0335
200	11	2,47	0,0189	10	0,0102
somme	592				0,9946



Distribution de fréquences relatives de la longueur (mm) de la rectrice centrale de 592 gélinottes huppées mâles à un moment donné superposée à la distribution des probabilités de chacune des classes selon la loi normale.

La loi de Student ou loi de *t*

La distribution de la variable *t* est utilisée pour comparer les moyennes de deux échantillons, tester une corrélation linéaire, la pente d'une régression, etc.
L'objectif ici sera d'apprendre à utiliser la table de *t*.

Description de la table

Les valeurs dans la table sont des valeurs de *t* ***et non*** des surfaces sous la courbe.
On note ces valeurs $t_{(\alpha:\upsilon)}$. La table ne donne que les valeurs positives car la distribution de *t* est **symétrique.**

- Les *valeurs* de la table sont des limites définies sur *l'abscisse* de la courbe.
- Les *probabilités* α ou α/2 (= les deux lignes d'en-tête du tableau)sont des *surfaces* sous la courbe.

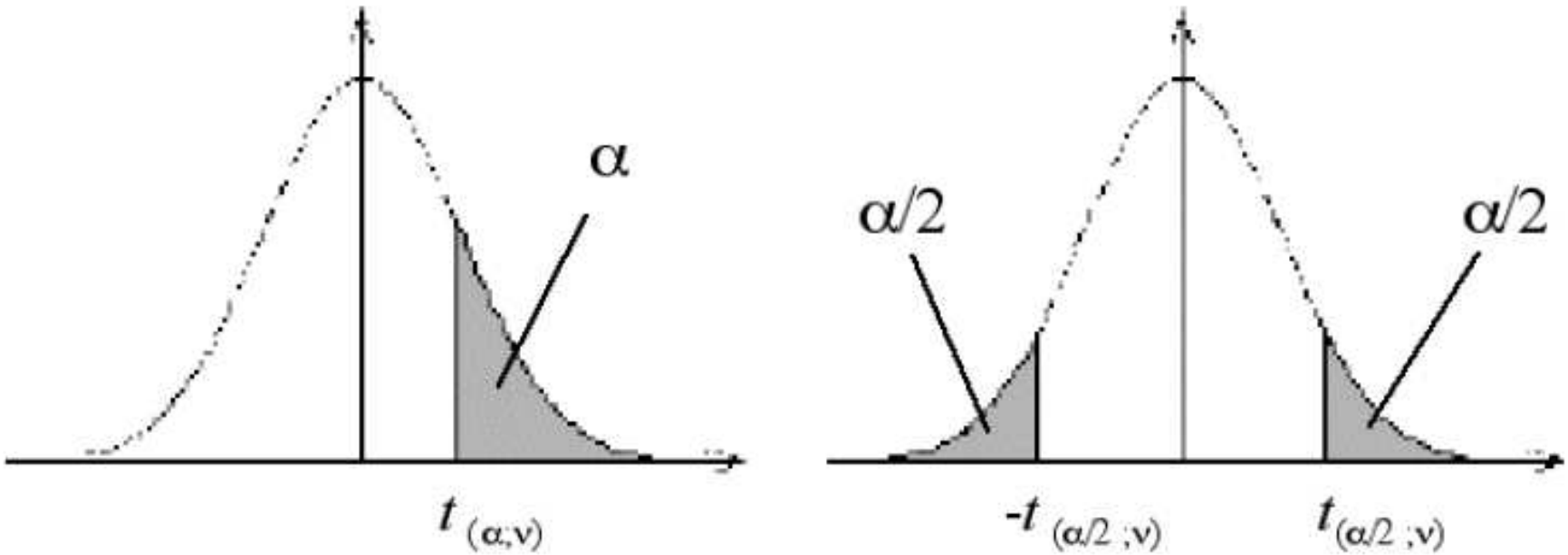
Valeurs critiques de la distribution du *t* de Student

α bilatéral	0.80	0.50	0.20	0.10	0.050	0.02	0.01	0.002	0.001
α unilatéral	0.40	0.25	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
ν									
1	0.325	1.000	3.078	6.314	12.71	31.82	63.66	318.3	636.6
2	0.289	0.816	1.886	2.920	4.303	6.695	9.925	22.33	31.60
3	0.277	0.765	1.638	2.353	3.182	4.541	5.841	10.21	12.92
4	0.271	0.741	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.267	0.727	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.265	0.718	1.440	1.943	2.447	3.143	3.707	5.208	5.959

La distribution change en fonction du nombre de *degrés de liberté* υ. Lorsque υ tend vers l'infini, la courbe de *t* converge vers une courbe normale centrée réduite.

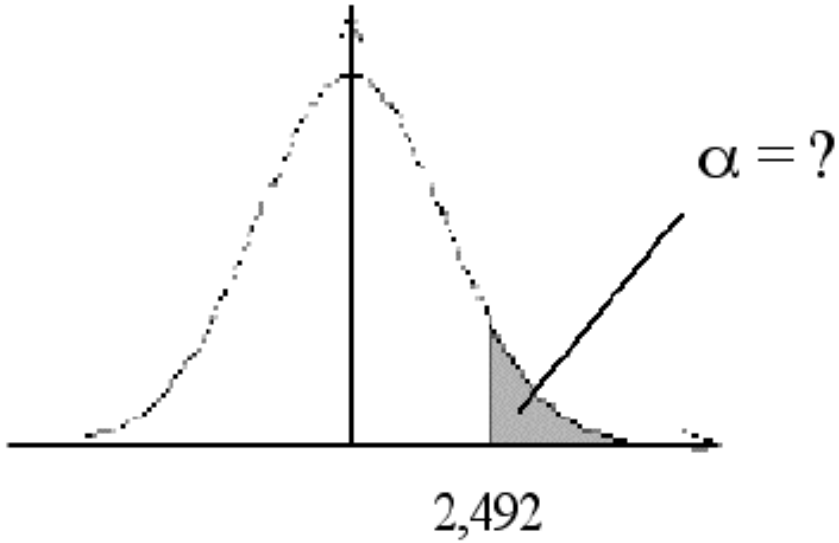
Le **seuil α** correspond à **P(*t* > *t* $_{(\alpha:\upsilon)}$)**, c'est-à-dire la probabilité que *t* égale ou dépasse une certaine *valeur critique*, définie en fonction du seuil de probabilité et du nombre de degrés de liberté.
Attention, le seuil peut être **unilatéral** ou **bilatéral**!!!

Si le seuil est bilatéral, la notation est la suivante: P($|t| > t_{(\alpha/2:\upsilon)}$)



Trouver la probabilité en connaissant les valeurs de $t_{(\alpha;\upsilon)}$

P(t_{24} > 2,492) = ? C'est-à-dire: quelle est la probabilité que la valeur de *t* pour 24 degrés de liberté soit **plus grande** que 2,492 ?



On voit que la probabilité est **unilatérale**

On lit dans la table *t*

$\nu = 24$

$t_{(\alpha;24)} = 2{,}492$

α unilatérale ----------> $\alpha = 0.01$

Pourquoi a-t-on pris la réponse à la ligne "α/2" de la table alors que la question est unilatérale ???

• Les valeurs données à la **ligne α** donnent la probabilité qu'une valeur de *t* soit située à **l'*extérieur de l'intervalle*** délimité par **[–*t* critique; +*t* critique].** Cet α est donc la ***somme*** des deux plages grises situées aux deux extrémités de la courbe.

• Les valeurs données à la **ligne α*/2*** donnent la probabilité qu'une valeur de *t* soit ***supérieure* au *t*** critique (si ce *t* est positif; plage grise de droite) ou ***inférieure* au *t*** critique (si ce *t* est négatif; plage grise de gauche).

• On peut donc interpréter la table de la manière suivante (avec notre exemple) :

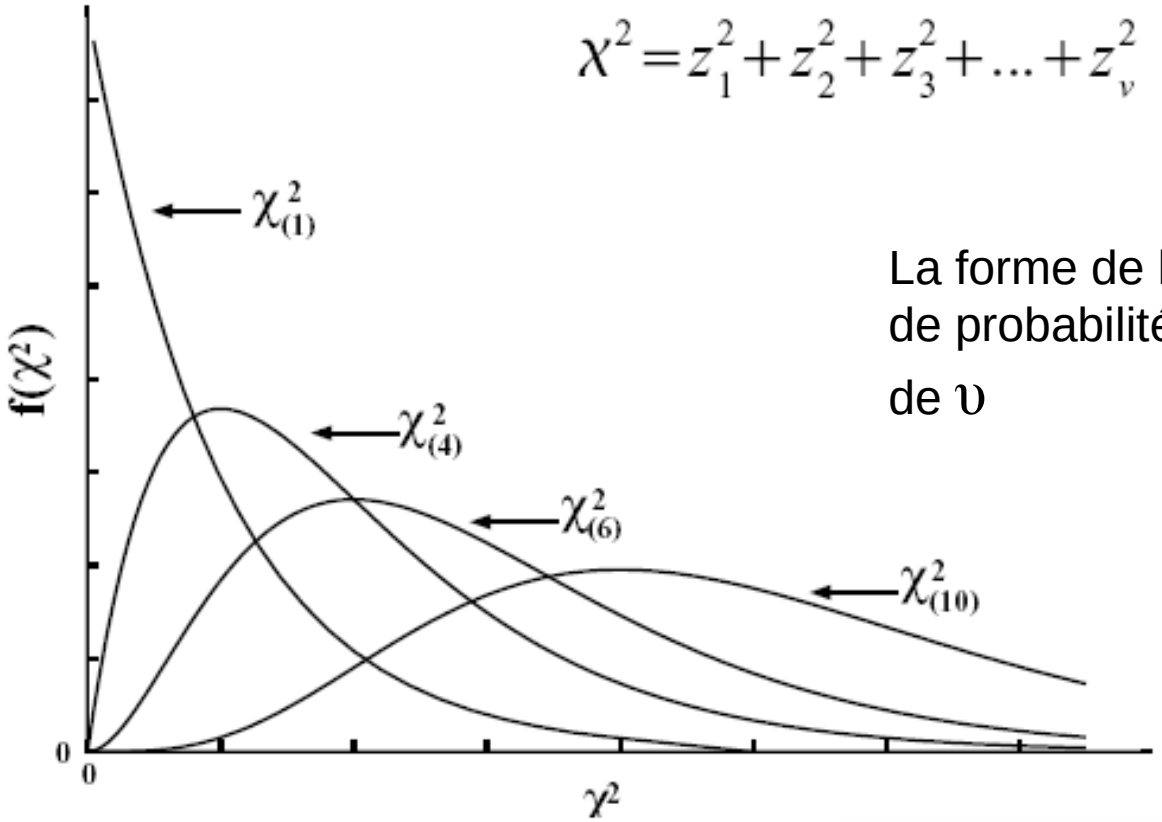
- si **la question est *unilatérale***, on veut connaître la probabilité qu'une valeur de ***t* soit *supérieure* au *t* critique**. Donc, on s'intéresse ***uniquement* à la valeur +2,492** (et non à –2,492). La surface située plus à droite que cette valeur limite représente 1% de la surface totale comprise sous la courbe exprimé α = 0,01;
- par contre, dans une **question *bilatérale***, formulée **P(|t_{24}| > 2,492),** on aurait voulu savoir quelle est la probabilité qu'une valeur de ***t* soit située à l'*extérieur* de l'intervalle** délimité par **[–2,492; +2,492].** Cette probabilité correspond à la somme des deux zones grises, soit α = 0.02 .

La loi du χ^2

Définition :

La loi de khi−carré est obtenue en faisant la somme des carrés de plusieurs lois normales :

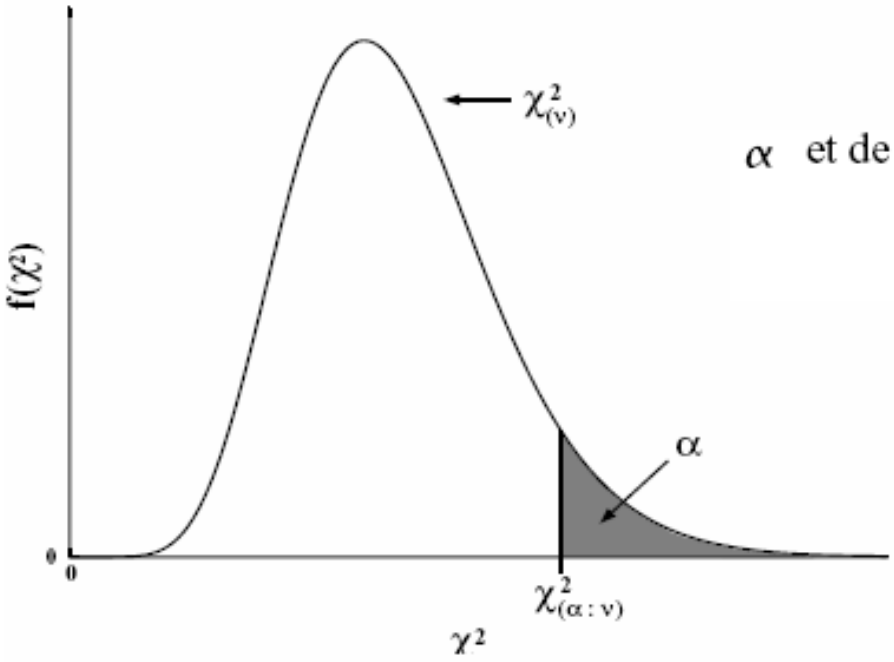
$$\chi^2 = z_1^2 + z_2^2 + z_3^2 + \ldots + z_v^2$$



La forme de la courbe de densité de probabilité change en fonction de υ

Distribution de densité des lois de χ^2 à $\nu=1$, $\nu=4$, $\nu=6$ et $\nu=10$ degrés de liberté.

Table de la loi de khi−carré $\chi^2_{(\alpha:\nu)}$



α et de nombre de degrés de liberté ν :

$$P\left(\chi^2_\nu > \chi^2_{(\alpha:\nu)}\right) = \alpha$$

La probabilité donnée dans la table est donc ***unilatérale à droite***.

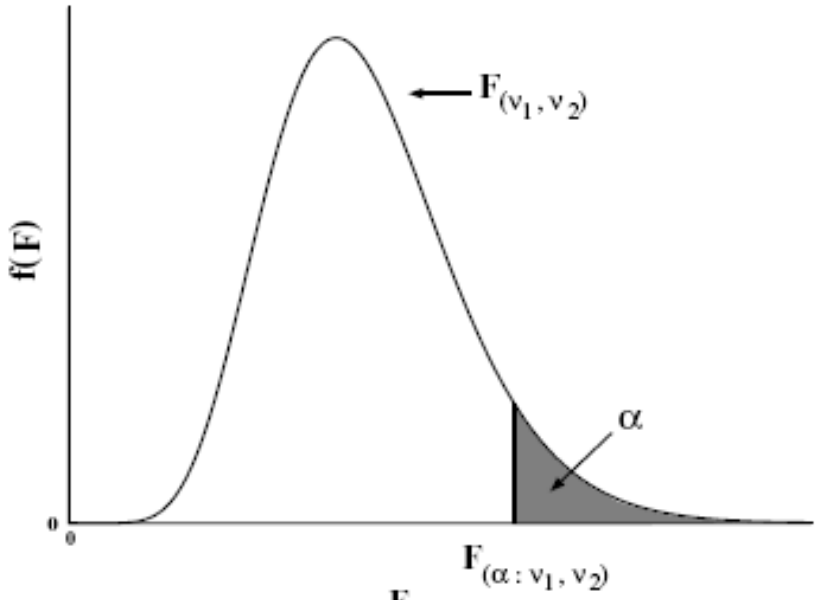
Loi de Fisher−Snedecor (*F*)

La loi de *F* est un rapport de deux lois de χ^2 à ν_1 et ν_2 degrés de liberté divisées par leur nombre respectif de degrés de liberté :

$$F_{(\nu_1,\nu_2)} = \frac{\chi^2_{\nu_1}/\nu_1}{\chi^2_{\nu_2}/\nu_2}$$

La forme de la courbe de densité de probabilité de *F* change en fonction de υ_1 et υ_2

Table de la loi de Fisher−Snedecor (*F*) $F_{(\alpha:\nu_1,\nu_2)}$



$$P\left(F_{(\nu_1,\nu_2)} > F_{(\alpha:\nu_1,\nu_2)}\right) = \alpha$$

La probabilité donnée dans la table est donc ***unilatérale à droite***.

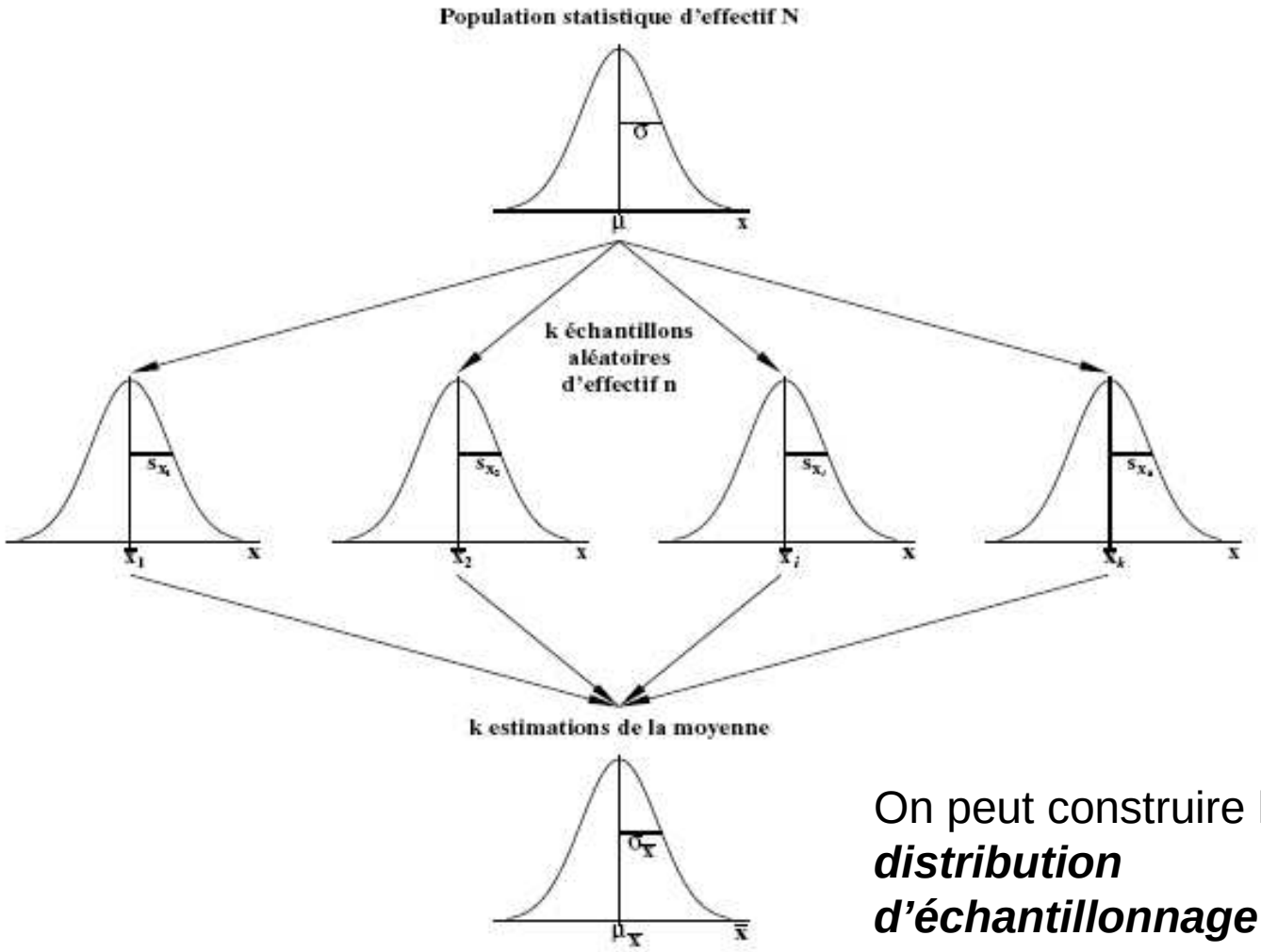
Intervalles de confiance (I.C.)

• Un I.C. d'un paramètre est une estimation ***par intervalle*** de ce paramètre.

Exemple : la moyenne de la variable *x* de la population statistique se situe entre telle et telle valeurs.

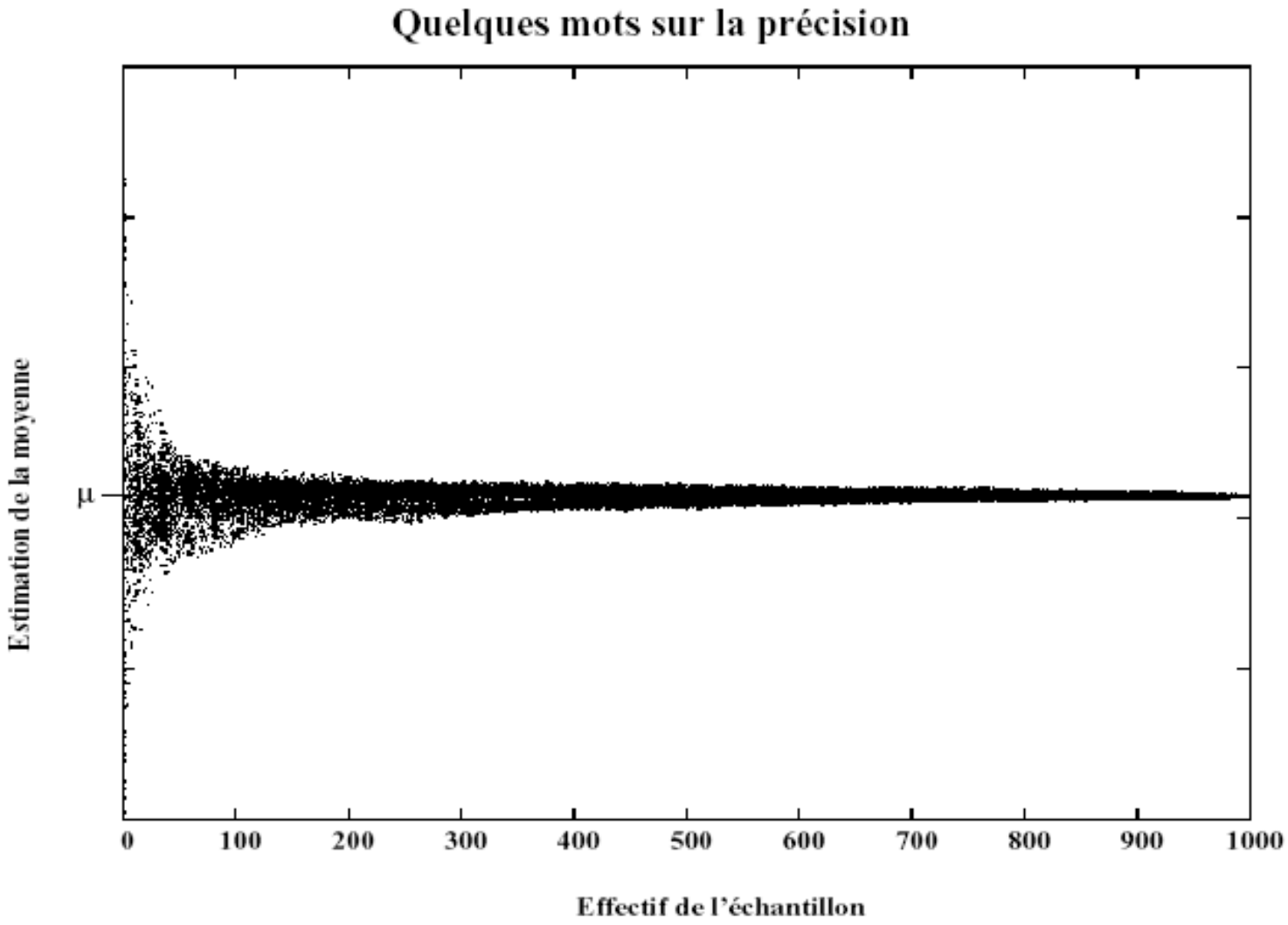
• On attribue un ***coefficient de risque*** (α) aux estimations par I.C.

Exemple : j'ai 5 % de chance de me tromper en disant que la moyenne de la variable *x* de la population statistique se situe entre telle et telle valeurs.
En général : α =0.05 *ou* 0.01 *ou* 0.001



On peut construire la ***distribution d'échantillonnage*** (ou de probabilités) de la moyenne.

On cherche alors les bornes de l'intervalle qui comprend 1-α des valeurs possibles de la moyenne selon sa distribution d'échantillonnage.



Intervalle de confiance de la moyenne : cas général

Dans ***tous les cas*** et surtout quand *n* est petit (*n*<30), la variable

$$t = \frac{\bar{x} - \mu}{s_{\bar{x}}}$$

obéit à une loi de Student avec $\nu = n - 1$ d.d.l.

L'I.C. de la moyenne peut alors être calculé comme suit :

$$P\left(\bar{x} - t_{\alpha/2:\nu} \cdot s_{\bar{x}} < \mu < \bar{x} + t_{\alpha/2:\nu} \cdot s_{\bar{x}} \right) = 1 - \alpha$$

- Variance de $\bar{x}$: $s_{\bar{x}}^2 = \frac{s_x^2}{n}$
- Erreur type de $\bar{x}$: $s_{\bar{x}} = \sqrt{s_{\bar{x}}^2} = \frac{s_x}{\sqrt{n}}$

Intervalle de confiance de la moyenne : cas des grands échantillons

Si *n* est très grands (*n*>30), la loi de *t* tends vers *N (*0,1) . On peut alors avoir recours une approximation par la loi normale :

$$P\left(\bar{x} - z_{\alpha/2:\nu} \cdot s_{\bar{x}} < \mu < \bar{x} + z_{\alpha/2:\nu} \cdot s_{\bar{x}} \right) = 1 - \alpha$$

ATTENTION : l'approximation par la loi normale n'est valable ***que*** si la distribution de l'échantillon n'est pas trop asymétrique! Il ***faut*** donc s'assurer que $n \geqslant 25 \cdot \alpha_3^2$ avant d'avoir recours à cette approximation.

Étant donné que le calcul de l'I.C. de la moyenne par la loi de *t* est ***toujours*** valable, on préférera cette méthode de calcul à l'approximation par la loi normale afin d'éviter les calculs supplémentaires et les inconvénients que cette dernière entraîne.

Moyenne et écart–type de la longeur totale de lézards mâles du genre *Cnemidophorus* provenant de trois populations du Nouveau Mexique (d'après Dessauer et al. 2000).

M : *C. t. marmotus*; P : *C. t. punctilinealis*; H : *C. t. marmotus* X *C. t. punctilinealis*

Site	**Espèce**	n	$\bar{x}$	s_x
14	M	61	86.9	7.13
26	H	19	84.4	7.06
20	P	22	82.0	6.28

En supposant que la distribution de la longueur totale obéit à une loi normale dans ces populations, calculez l'I.C. à 95 % de la moyenne pour chacune de ces populations.

$$P\left(\bar{x} - t_{\alpha/2:\nu} \cdot s_{\bar{x}} < \mu < \bar{x} + t_{\alpha/2:\nu} \cdot s_{\bar{x}} \right) = 0.95$$

- **Population 14 : *C. t. marmotus***

$n = 61$ donc $\nu = n - 1 = 60$ et $t_{\alpha/2:\nu} = t_{0.025:60} = 2.000$

$$P\left(86.9 - 2.000 \cdot \frac{7.13}{\sqrt{61}} < \mu < 86.9 + 2.000 \cdot \frac{7.13}{\sqrt{61}} \right) = 0.95$$

$$P\left(85.07419... < \mu < 88.72581... \right) = 0.95$$

$$I.C.\,95\%\,\mu = [85.1\,mm, 88.7\,mm]$$

- **Population 26 : *C. t. marmotus* X *C. t. punctilinealis***

$t_{\alpha/2:\nu} = t_{0.025:18} = 2.101$ $\quad I.C.\,95\%\,\mu = [81.0\,mm, 87.8\,mm]$

- **Population 20 : *C. t. punctilinealis***

$t_{\alpha/2:\nu} = t_{0.025:21} = 2.080$ $\quad I.C.\,95\%\,\mu = [79.2\,mm, 84.8\,mm]$

Intervalle de confiance de la variance

Si les observations proviennent d'un échantillon aléatoire simple extrait d'une seule population statistique, la distribution de probabilités de la variance suit une loi de

$$\chi^2_\nu = \frac{\sum (x-\bar{x})^2}{\sigma^2} \quad \text{à} \quad \nu = n-1 \ \text{ d.d.l.}$$

Cette formule peut être réécrite sous la forme :

$$\chi^2_\nu = \frac{(n-1)\cdot s_x^2}{\sigma^2}$$

L'I.C. de la variance est donc :

$$P\left(\frac{(n-1)\cdot s_x^2}{\chi^2_{\alpha/2}} < \sigma^2 < \frac{(n-1)\cdot s_x^2}{\chi^2_{1-\alpha/2}} \right) = 1-\alpha$$

Intervalle de confiance de l'écart type

$$P\left(\sqrt{\frac{(n-1)\cdot s_x^2}{\chi^2_{\alpha/2}}} < \sigma < \sqrt{\frac{(n-1)\cdot s_x^2}{\chi^2_{1-\alpha/2}}} \right) = 1-\alpha$$