TD de Statistiques - Séance N°4
Lois théoriques - Tests statistiques

1 Lois théoriques

1.1 Lois théoriques discrètes

On se donne un ensemble de modalités (par exemple : 0, 1, ..., *N*) et une formule mathématique permettant de calculer leurs fréquences d'apparition.

Pour que la loi ait un intérêt en pratique, il faut qu'elle puisse servir à modéliser des situations du monde réel.

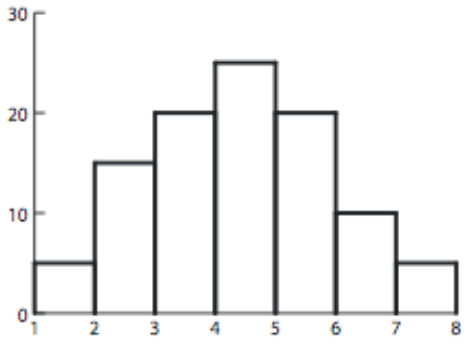
Exemple 1 : Une épreuve élémentaire, avec seulement deux issues possibles (succès, échec) est répétée un nombre déterminé de fois, *N*, de façon indépendante. Les chances de succès à chaque épreuve sont égales à p ($0 \leq p \leq 1$). On compte le nombre de succès observés sur les *N* épreuves : il s'agit alors de la loi binomiale de paramètres *N* et p.

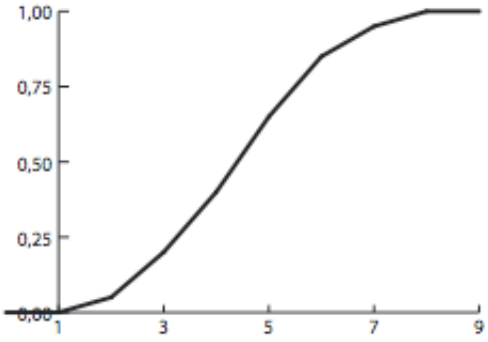
Exemple 2 : Une épreuve élémentaire, avec seulement deux issues possibles (succès, échec) est répétée jusqu'à l'obtention du premier succès. Les chances de succès à chaque épreuve sont égales à p ($0 \leq p \leq 1$). On compte le nombre d'échecs rencontrés avant l'obtention du premier succès : c'est la loi géométrique de paramètre p.

1.2 Lois théoriques continues

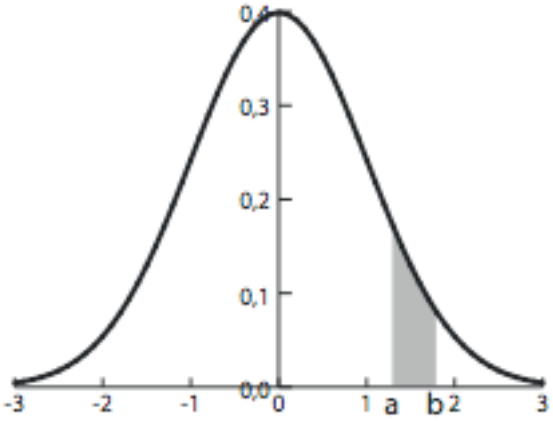
Du point de vue mathématique, on se donne une fonction *f* telle que l'aire située sous la courbe $y=f(x)$ soit égale à 1. *f* est appelée densité de la loi.

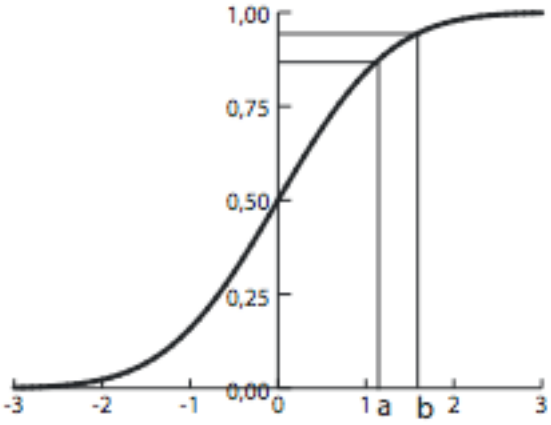
La densité d'une loi théorique est la modélisation mathématique de la notion d'histogramme pour une distribution empirique.





De façon analogue, une loi théorique de distribution statistique est donnée par sa densité $f(x)$ ou sa fonction de répartition : $F(x)$.





La fréquence (le pourcentage d'observations) vérifiant $a \leq X \leq b$ est donnée par l'aire hachurée ou par la valeur $F(b)$ - $F(a)$.

1.3 Loi Normale ou loi de Laplace Gauss

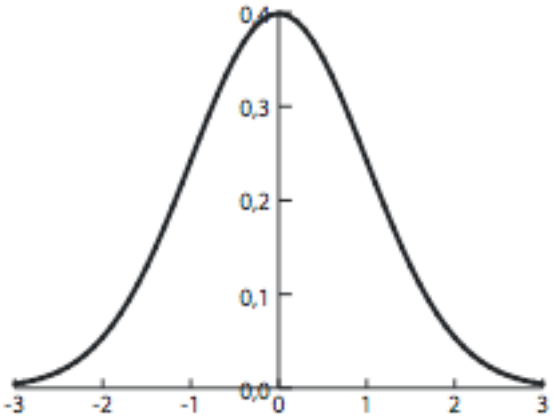
Problème : trouver une loi théorique modélisant la distribution d'une variable numérique dont les valeurs résultent d'une combinaison d'effets *nombreux, indépendants entre eux, additifs et de même ordre de grandeur*.

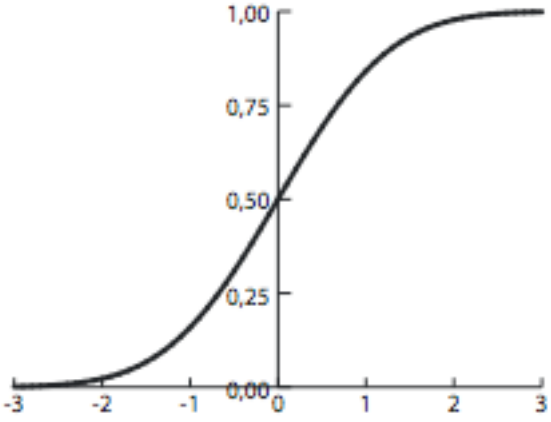
Réponse : La loi normale.

1.3.1 Loi normale centrée réduite

Moyenne : $\mu = 0$. Ecart type : $\sigma = 1$

Densité : $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$





1.3.2 Loi normale, cas général : transformation en *Z*

La variable *X* suit une loi normale de paramètres μ et σ si la variable *Z* définie par :

$$Z = \frac{X - \mu}{\sigma}$$

suit une loi normale centrée réduite.

La transformation précédente est appelée "transformation en *Z*" ou "centrage réduction" de la variable.

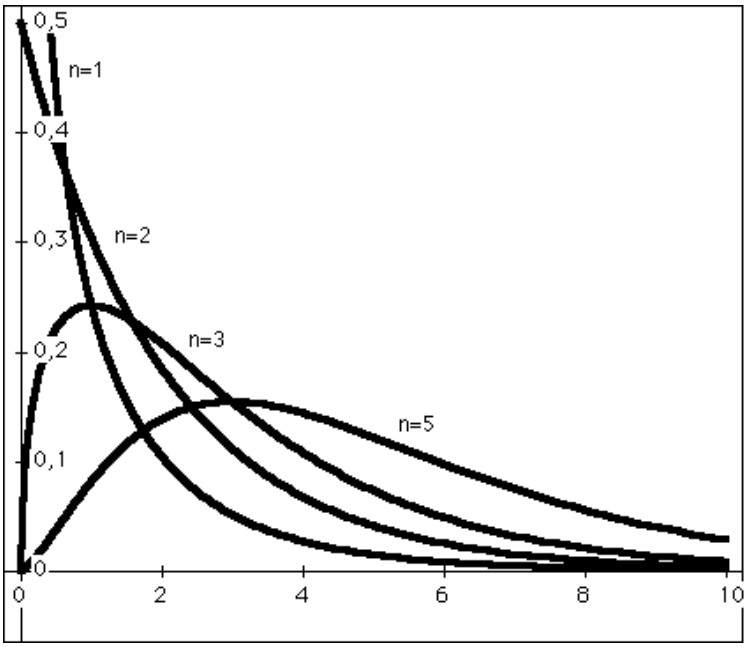
Il existe une loi analogue, permettant une meilleure modélisation de certaines situations : *la loi de Student à n degrés de liberté*.

1.4 Loi du khi-2

Problème : Dans une situation donnée, on considère *n* variables normales centrées réduites indépendantes. On s'intéresse à la somme des carrés de ces variables : $S = X_1^2 + X_2^2 + ... + X_n^2$.

La loi suivie par *S* est la loi du χ^2 (khi-2, chi-2) à *n* degrés de liberté.

Distributions du khi-2 :



Il existe une loi analogue, permettant une meilleure modélisation de certaines situations : la loi de Fisher-Snedecor à n_1 et n_2 degrés de liberté}

2 Tests statistiques

2.1 Notion de test statistique

Comment séparer le probable de l'improbable ?

Exemple intuitif. Hier soir, un match de foot était retransmis à la télévision. Le score final a été : 14 à 1. Qu'en pensez-vous?

Deux hypothèses :
H_0 : L'affirmation est correcte.

H_1 : L'affirmation est erronée. Par exemple, le score annoncé est incorrect, ou alors, il ne s'agissait pas de football mais de rugby ou de handball...

Difficile de raisonner sur H_1 (trop imprécise, pas d'information suffisante). Raisonnons en supposant H_0 vraie.

Les règles du jeu de football autorisent tout à fait un score tel que 14 à 1. Mais, nous avons une connaissance (intuitive) de la distribution des scores des matches de foot et nous savons qu'un tel score est extrêmement rare (aussi bien que des scores plus extrêmes : 15 à 0, 15 à 1, etc).

Entre les deux explications :

- le match a abouti à un score tout à fait exceptionnel ;
- l'affirmation est incorrecte

nous choisissons plutôt la seconde : l'affirmation est incorrecte, car la première explication est trop improbable.

Mais, pour un score de 4 à 1, nous aurions fait le choix inverse, et pour un score de 6 à 1, nous n'aurions pas trop su comment conclure...

2.2 Observer un échantillon

Exemple (idiot) : on veut évaluer la taille moyenne de la population française adulte.

Solution 1. On interroge de façon exhaustive le fichier des cartes d'identité. Mais c'est très coûteux, et la CNIL proteste ...

Solution 2. On tire au hasard 1 sujet. Il mesure 174 cm. Mais cela ne fournit pas beaucoup d'information. Notamment, on n'a aucune indication sur la variabilité de la grandeur observée : les tailles s'échelonnent-elles de 80 à 250 ou de 170 à 180 ?

Solution 3. On tire au hasard, avec remise, $n = 100$ sujets dans la population française adulte. On peut en tirer de multiples informations :

- La moyenne $\bar{x}$ des tailles observées sur l'échantillon me fournit un *estimateur non biaisé* de la taille moyenne sur la population.

- L'écart type et la variance observés sur l'échantillon (appelés écart type et variance *empiriques*) fournissent des estimateurs biaisés de ces paramètres dans la population. Pour supprimer ce biais :

$$\begin{bmatrix}\text{Variance estimée}\\ \text{dans la population}\end{bmatrix} = \frac{n}{n-1}\begin{bmatrix}\text{Variance observée}\\ \text{sur l'échantillon}\end{bmatrix}$$

Cette variance estimée est appelée *variance corrigée* ou *variance estimée*.

- On a une idée de la "précision" de cet estimateur de la taille moyenne. C'est une variable statistique distribuée (presque) normalement, dont la moyenne est la taille moyenne de la population et dont la variance est celle de la population divisée par l'effectif ($n = 100$) de l'échantillon. Autrement dit, cet estimateur est d'autant plus "précis" que la taille de l'échantillon est grande.

2.3 Indépendance de deux variables nominales - Test du khi-2

2.3.1 Indépendance de deux variables nominales

Deux variables nominales X et Y sont observées sur un échantillon de sujets.
Nombre de modalités de X : l
Nombre de modalités de Y : c.

Problème : ces deux variables sont-elles indépendantes entre elles ?

Exemple : On a observé trois groupes de musiciens : musiciens professionnels (MP), musiciens en cours de professionnalisation (MCP) et musiciens amateurs (MA).
On s'intéresse au niveau d'études des trois groupes.

Les effectifs observés sont donnés par le tableau de contingence suivant :

	MP	MCP	MA	Total
Avant bac.	7	11	4	22
bac.	12	6	5	23
Post bac.	17	13	20	50
Total	36	30	29	95

Le niveau d'études et type de professionnalisation sont-ils liés ?

Fréquences lignes : en l'absence de lien, les fréquences sur chaque ligne devraient être proches des fréquences de la ligne "Synthèse" :

	MP	MCP	MA	Total
avant bac.	32%	50%	18%	100%
bac.	52%	26%	22%	100%
post bac.	34%	26%	40%	100%
Synthèse	38%	32%	31%	100%

Fréquences colonnes : de même, en l'absence de lien, les fréquences dans chaque colonne devraient être proches des fréquences de la colonne "Synthèse" :

	MP	MCP	MA	Synthèse
avant bac.	19%	37%	14%	23%
bac.	33%	20%	17%	24%
post bac.	47%	43%	69%	53%
Total	100%	100%	100%	100%

Mais nos observations portent sur un échantillon. Les différences constatées peuvent-elles être expliquées par les *fluctuations d'échantillonnage* ?

2.3.2 Test proprement dit

Hypothèses :

H_0 : Les variables X et Y sont indépendantes.
H_1 : Les variables X et Y sont dépendantes.

Statistique de test

Une statistique est une variable dont on peut calculer la valeur sur chaque échantillon susceptible d'être tiré et dont la loi de distribution est une loi théorique connue. Pour la situation envisagée, la statistique à utiliser est la *distance du khi-2* entre le tableau des effectifs observés et un tableau *d'effectifs théoriques* (cf. calcul infra).

Si l'hypothèse H_0 est vérifiée (autrement dit, sous l'hypothèse H_0), cette statistique suit une loi du khi-2 à $(l\text{-}1)(c\text{-}1)$ degrés de liberté (ddl).

Calcul de la distance du khi-2

Données observées : tableau de contingence.

Effectifs attendus (ou théoriques) sous hypothèse d'indépendance :

Dans chaque case : $\text{Effectif théorique} = \dfrac{\text{total ligne} \times \text{total colonne}}{\text{total général}}$

Contribution de chaque case au khi-2 : $\text{Ctr}_{ij} = \dfrac{(\text{Effectif observé} - \text{Effectif théorique})^2}{\text{Effectif théorique}}$

Distance du khi-2 : $\chi^2 = \sum_{ij} \text{Ctr}_{ij}$.

Calcul sur l'exemple traité :

Rappel des effectifs observés

	MP	MCP	MA	Total
Avant bac.	7	11	4	22
bac.	12	6	5	23
Post bac.	17	13	20	50
Total	36	30	29	95

Effectifs théoriques

	MP	MCP	MA
Avant bac.	8,34	6,95	6,72
bac.	8,71	7,26	7,02
Post bac.	18,95	15,79	15,26

Exemple de calcul d'un effectif théorique : $\dfrac{22 \times 36}{95} = 8{,}34$.

Calcul de la distance du khi-2 :

Modalités	n_{ij}	t_{ij}	$\dfrac{(n_{ij} - t_{ij})^2}{t_{ij}}$
MP. < Bac	7	8.34	0,21
MP. Bac	...		1,23
MP. > Bac			0,20
MCP. < Bac			2,36

MCP. Bac			0,22
MCP. > Bac			0,49
MA. < Bac			1,09
MA. Bac			0,58
MA. > Bac			1,47
Total			7,85

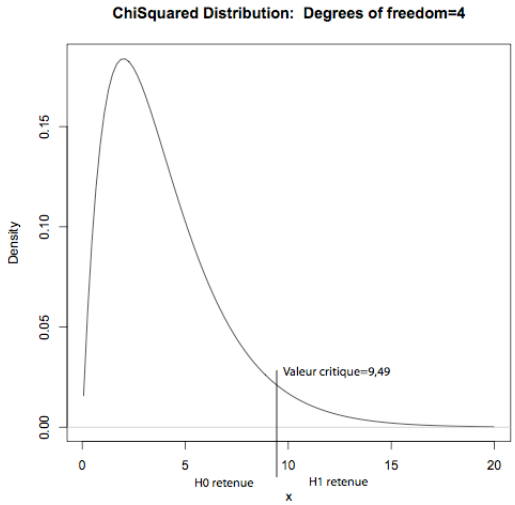
On obtient : $\chi^2_{obs} = 7{,}85$.

La suite de la démarche diffère selon qu'il s'agit d'un calcul à la main ou d'un résultat produit par un logiciel de traitement statistique.

Règle de décision et conclusion en calcul manuel :

- On choisit un seuil, par exemple $\alpha = 5\%$.
- Le nombre de degrés de liberté est ici : $(3-1)(3-1) = 4$.
- On lit dans la table la valeur critique du khi-2 à 4 ddl pour un seuil de 5% : $\chi^2_{crit} = 9{,}49$.

L'hypothèse H_0 correspond aux petites valeurs du khi-2 (cf. le mode de la distribution ci-dessous), pendant que H_1 sera retenue pour les valeurs élevées du khi-2 :



On formule donc la règle de décision suivante :

- Si $\chi^2_{obs} \leq \chi^2_{crit}$, on retient H_0.
- Si $\chi^2_{obs} > \chi^2_{crit}$, on retient H_1.

Sur notre exemple : $\chi^2_{obs} = 7{,}85$, $\chi^2_{crit} = 9{,}49$ et donc $\chi^2_{obs} \leq \chi^2_{crit}$. On retient donc H_0 : on n'a pas mis en évidence de différence de niveau d'étude selon le type de professionnalisation.

Règle de décision et conclusion lorsqu'on utilise un logiciel de traitement statistique :

Le logiciel nous indique, pour une loi du khi-2 à 4 ddl, la probabilité d'observer une valeur supérieure ou égale à χ^2_{obs}. C'est le *niveau de significativité* ou *p-value* du test.
Sur notre exemple : p - value = 0,097 = 9,7%.

Cette valeur est ici supérieure aux seuils traditionnels (5%, 1%, ...). Comme précédemment, on conclut sur H_0.

2.3.3 Remarques

Remarques sur le test du khi-2

1. Condition sur les effectifs théoriques minimaux :
Le test du khi-2 ne peut pas être appliqué si les effectifs sont trop faibles. On exige en général :
- que, dans le tableau des effectifs théoriques, les effectifs strictement inférieurs à 5 représentent moins de 20% des cases;
- et que, dans le tableau des effectifs théoriques, ne figure aucun effectif inférieur à 1.

2. Dans le cas d'un tableau à deux lignes et deux colonnes, on opère souvent une correction au calcul du khi-2 : la correction de Yates. Cette correction a un effet d'autant plus important que les effectifs sont faibles.

3. Dans le cas d'un tableau à deux lignes et deux colonnes, si les effectifs sont trop faibles pour utiliser le test du khi-2, on peut utiliser le *test exact de Fisher*.

4. Le test du khi-2 peut notamment servir à comparer deux proportions (cf. TD avec R).

Remarque sur les tests statistiques en général : seuil et p-value.

La règle de décision par valeur observée et valeur critique et celle formulée à l'aide d'une p-value sont liées par les propriétés suivantes :

- Si la p-value est inférieure au seuil α qui a été fixé, c'est l'hypothèse H_1 qui doit être retenue.

- Si la p-value est supérieure ou égale au seuil α qui a été fixé, c'est l'hypothèse H_0 qui doit être retenue.

2.4 Test de significativité d'un coefficient de corrélation

2.4.1 Principe du test

Deux variables numériques X et Y sont observées sur un échantillon de n sujets. Soit r leur coefficient de corrélation.

- Les données (x_i, y_i) constituent un échantillon.
- Le coefficient r est une statistique.

On introduit le coefficient ρ : coefficient de corrélation inconnu dans la population.

Les hypothèses du test du coefficient de corrélation s'écrivent :

H_0 : Indépendance de X et Y sur la population, c'est-à-dire : $\rho = 0$.
H_1 : $\rho \neq 0$ (test bilatéral)

Statistique de test

- Pour de petits échantillons, des tables spécifiques donnent directement les valeurs critiques de r. Le nombre de degrés de liberté à prendre en compte est $ddl = n - 2$.

- De manière alternative et pour de grands échantillons, on calcule : $T = \sqrt{n-2}\,\frac{r}{\sqrt{1-r^2}}$.

Sous H_0, T suit une loi de Student à n - 2 degrés de liberté.

Conditions d'application

Dans la population parente, le couple (X, Y) est supposé suivre une loi normale bivariée, ce qui implique notamment :

- la normalité des distributions marginales de X et Y ;
- la normalité de la distribution de l'une des variables lorsque l'autre variable est fixée ;
- l'égalité des variances des distributions de l'une des variables pour deux valeurs distinctes de l'autre variable.

2.4.2 Exemple :

Source : Article disponible en ligne à l'adresse: http://www.cairn.info/revue-deviance-et-societe-2012-1-page-3.htm

On a mené une étude auprès de 26 collèges publics du département du Nord. 9 d'entre eux étaient classés en ZEP tandis que 17 autres ne possédaient aucune qualification particulière. On dispose, notamment, pour chaque établissement d'une mesure du niveau de violence ressenti par les élèves et par les professionnels et d'une mesure du taux d'encadrement.

Pour les 9 établissements classés en ZEP, le coefficient de corrélation entre la violence ressentie par les élèves et le taux d'encadrement est $r_1 = 0{,}62$, tandis que le coefficient de corrélation entre la violence ressentie par les professionnels et le taux d'encadrement est $r_2 = 0{,}72$. Ces coefficients sont-ils significatifs d'un lien entre les deux variables invoquées, au seuil de 5% ?

Réponse : pour $ddl = 7$ et un seuil de 5%, on lit dans la table $r_{crit} = 0{,}6664$. On n'a pas mis en évidence de lien entre les deux variables dans le cas des élèves (mais la taille de l'échantillon est faible). En revanche, on conclut à l'existence d'un lien entre les deux variables dans le cas des professionnels.

Etudier de même la situation des 17 établissements "ordinaires", sachant que les deux coefficients de corrélation sont alors $r_1 = 0{,}35$ et $r_2 = 0{,}68$.

Réponse : pour $ddl = 15$ et un seuil de 5%, on lit dans la table $r_{crit} = 0{,}4821$. Comme précédemment, on n'a pas mis en évidence de lien entre les deux variables dans le cas des élèves; en revanche, on conclut à l'existence d'un lien entre les deux variables dans le cas des professionnels.

2.5 Tests de comparaison de moyennes

2.5.1 Comparaison de deux moyennes sur des groupes indépendants

Une variable X est observée sur deux échantillons tirés au hasard dans deux populations différentes. Au vu de ce qui est observé sur les deux échantillons, les moyennes de X dans les populations parentes sont-elles égales ou différentes ? Ou encore, les moyennes observées sur ces deux échantillons sont-elles significativement différentes ?

Notations

μ_1, μ_2 : moyennes *inconnues* sur les populations parentes respectives (condition d'application : distributions normales de même variance).

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

Hypothèses du test

H_0 : $\mu_1 = \mu_2$
H_1 : (test bilatéral) $\mu_1 \neq \mu_2$

Statistique de test

$$T = \frac{\text{Différence des moyennes observées dans les deux groupes}}{\text{Erreur type}}$$

L'erreur type est calculée à partir des écarts types observés dans les deux groupes et des effectifs des échantillons.

Sous H_0, T suit la loi de Student à $n_1 + n_2 - 2$ *ddl*. La loi de Student peut être assimilée à une loi normale centrée réduite si $n_1 > 30$ et $n_2 > 30$.

2.5.2 Exemple :

On observe un groupe de 30 adultes jeunes, et un groupe de 30 adultes âgés. On soumet les sujets des deux groupes à une épreuve de fluence orthographique. Les paramètres calculés à partir des résultats observés sont les suivants :

	Jeunes	Agés
n	30	30
$\bar{x}$	11,4	11,0
s_c	3,1	3,2

H_0 : $\mu_1 = \mu_2$
H_1 : (test bilatéral) $\mu_1 \neq \mu_2$

La statistique de test suit une loi de Student à 30 + 30 - 2 = 58 *ddl*.

Pour un seuil de 5%, la valeur critique lue dans la table est $t_{crit} = 2.0017$. La règle de décision est donc :

- si $-2{,}0017 \leq t_{obs} \leq 2{,}0017$, on retient H_0.
- si $t_{obs} < -2.0017$ ou $t_{obs} > 2.0017$, on rejette H_0 et on retient H_1.

Or (calcul réalisé à l'aide d'un logiciel) : $t_{obs} = \frac{11{,}4 - 11{,}0}{0{,}81} = 0{,}4917$

On retient donc l'hypothèse H_0 : on n'a pas mis en évidence de différence significative de la fluence verbale.

2.5.3 Comparaison de deux moyennes. Groupes appariés

Une variable *X* est observée, dans deux conditions différentes, sur un échantillon tiré au hasard dans une population. Au vu de ce qui est observé, les moyennes de *X* sur la population parente dans les deux conditions sont-elles égales ou différentes ? Ou encore, les moyennes observées dans ces deux conditions sont-elles significativement différentes ?

On introduit le protocole dérivé des différences individuelles ($d_i = x_{i1} - x_{i2}$)

Notations

μ_1, μ_2 : moyennes *inconnues* de la variable *X* sur la population parente, dans les deux conditions.
δ : 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 sur des échantillons de tailles n_1 et n_2
$\bar{d}$: 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 : $\mu_1 = \mu_2$
H_1 : (test bilatéral) $\mu_1 \neq \mu_2$

Statistique de test.

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

Sous H_0, T suit la loi de Student à n - 1 *ddl*.

Pour $n > 30$, la loi de Student peut être assimilée à une loi normale centrée réduite.

2.5.4 Exemple

On a mesuré le temps de réaction de 10 sujets à jeun ($\bar{x}_1 = 22{,}3$ ms) et sous l'influence d'un tranquilisant ($\bar{x}_2 = 31{,}7$ ms). L'écart type corrigé de la série des différences est : $s_c = 11{,}54$.

H_0 : $\delta = 0$
H_1 : $\delta \neq 0$ (test bilatéral, par exemple).

La statistique de test *T* suit une loi de Student à 9 *ddl*, et, pour un seuil de 5%, la valeur critique est : $t_{crit} = 2{,}26$.

Or : $E^2 = \frac{11{,}54^2}{10} = 13{,}32$, $E = 3{,}65$, $T_{obs} = \frac{22.3 - 31{,}7}{3{,}65} = -2{,}58$

On conclut donc sur H_1.