Statistiques déductives :

I Généralités sur les test d'hypothèse :

Dans les statistiques déductives, contrairement aux statistiques descriptives, on essaie, à partir des observations faites, de tirer des conclusions. Pour cela, les épidémiologistes utilisent des tests d'hypothèse.

A Les tests de comparaison :

Le plus souvent, les test utilisés en statistiques déductives sont des **tests de comparaison** soit :

- entre 2 populations : on constitue alors 2 échantillons représentatifs et on essaie de déterminer *s'il existe une différence significative entre ces 2 échantillons pour le caractère étudié*. Le but étant d'extrapoler le résultat aux 2 populations primitives.

- Entre une population donnée A et la population générale de référence : on constitue alors un échantillon représentatif de la population A et on essaie de déterminer *s'il existe une différence significative entre l'échantillon et la population générale* et donc, in fine, entre A et la population générale.

B La définition des hypothèses :

La première étape d'une étude statistique, ici d'un test de comparaison, est la **formulation d'hypothèses** que le test permettra ensuite de confirmer / infirmer. On définira, au début de chaque test, 2 hypothèses jouant un rôle symétrique :

1. **H0= hypothèse nulle** :
   - « Il n'y a pas de différence observée entre les deux groupes »
   - « Il n'existe pas de lien entre les 2 caractères étudiés, les fluctuations observées sont donc dues au hasard»

2. **H1 = hypothèse alternative**.
   - « Il y a une différence significative entre les deux groupes »
   - « Il existe bien un lien entre les 2 caractères étudiés, les fluctuations observées ne sont donc pas dues au hasard. »

Les tests sont donc des techniques permettant de décider si on accepte ou si on rejette H0, en ayant fixé le risque d'erreur α accompagnant cette décision.

NB : On choisit toujours pour **H0** l'hypothèse qu'il serait le plus grave de rejeter à tort.

C La notion de risque :

Rappel de statistique descriptive : Lors de l'estimation d'une valeur x par un IC, α représente le **risque d'erreur** dans l'estimation de x, c'est à dire *le risque pour que l'IC ne contienne pas la vraie valeur de x*. Il est généralement fixé à **5%**

En statistique déductive, on a :

1. α ou risque de première espèce représente : **le risque de rejeter H0 si H0 est vraie**. Ce risque d'erreur est maîtrisé, c'est à dire qu'il est fixé (le plus souvent à 5%) avant l'application du test statistique.

2. 1 – α représente la **probabilité d'accepter H0 si H0 est vrai**

3. β ou risque de seconde espèce représente le **risque d'accepter H0 si H0 est fausse**. Ce risque d'erreur est négligé et peut donc être assez important.

4. 1- β représente la **puissance du test**. Il s'agit de la **probabilité de rejeter H0 si H0 est fausse.**

On définit la **règle du rejet** uniquement à partir de **H0** et **d'α**. Le tableau suivant récapitule bien toutes ces notions :

		α	1-α	β	1-β
H0	**vraie/fausse ?**	vraie	vraie	fausse	fausse
	Acceptation / rejet?	rejet	acceptation	acceptation	rejet
H1	**Vraie / fausse ?**	fausse	fausse	vraie	vraie
	Acceptation/rejet ?	acceptation	rejet	rejet	acceptation

D Les étapes d'un test d'hypothèse :

Pour mettre en oeuvre un test d'hypothèse, on suivra TOUJOURS les étapes suivantes :

1. définir H0 et H1. Les deux hypothèses jouent des rôles **symétriques**.

2. Déterminer le caractère des données à étudier/comparer :
   - qualitative/qualitative
   - qualitative/quantitative
   - quantitative/quantitative

3. Choisir le test en fonction du type de données On nomme **Z** le paramètre qui sera calculé.

4. Choisir le seuil d'erreur de $1^{ère}$ espèce α généralement fixé à **5%**.

5. Recueillir les données, calculer Z et utiliser la règle de rejet (définie à partir de H0 et de α) : Il s'agit de **comparer Z par rapport à une valeur théorique** de référence

6. Interpréter des résultats :
   - Au niveau de l'échantillon : Accepte -t-on H0 ?
   - Au niveau de la population : peut-on extrapoler les résultats à la population générale ?

NB : L'acceptation de H0 implique forcément le rejet de H1 et *vice versa*.

II L'étude de la liaison entre deux caractères qualitatifs :

Dans la suite du cours, nous garderons, pour tous les types de test abordés, la méthode d'application d'un test d'hypothèse mise en place ci-dessus. Pour l'étude de la liaison entre 2 caractères qualitatifs, on peut choisir d'utiliser soit :

1. un test de comparaison des pourcentages
2. un test du X^2

Toutes les formules (sauf celles des ddl) fournies dans la suite de la fiche ne sont pas à connaître !

II L'étude de la liaison entre deux caractères qualitatifs :

A Le test de comparaison des pourcentages :

Soient **2 groupes A et B** et une caractéristique **qualitative x** (couleur des yeux etc.) . On peut se demander si la proportion d'individus du goupe A présentant x coïncide avec la proportion d'individus du goupe B présentant x.

1) définir H0 et H1.

- **H0** : il n'y a pas de différence observée entre les groupes A et B, c'est à dire que la la proportion d'individus du goupe A présentant x coïncide avec la proportion d'individus du goupe B présentant x.

- **H1** : il existe une différence signitificative entre les groupes A et B, c'est à dire que la la proportion d'individus du goupe A présentant x est différente de celle du goupe B.

2) Déterminer le caractère des données à étudier/comparer : Les variables sont :
- des types d'individus → variable **qualitative**
- une caractéristique x **qualitative**

3) Choisir le test en fonction du type de données : En présence de données qualitatives, on peut choisir d'utiliser un **test de comparaison des pourcentages**.

4) Choisir le seuil d'erreur de 1ère espèce α généralement fixé à **5%.**

5) Recueillir les données, calculer Z et utiliser la règle de rejet
La variable Z est ici représentée par **l'écart réduit ε**. On comparera :
1. εt_h donné par la table de l'écart réduit, en fonction de α
2. $\varepsilon_{calculée} = \varepsilon_{exp} = \dfrac{p_A - p_B}{\sqrt{\dfrac{p_A q_A}{n_A} + \dfrac{p_B q_B}{n_B}}}$

6) Interpréter les résultats : Au niveau de l'échantillon :
1. si $\varepsilon_{calculée} > \varepsilon t_h$ alors **on rejette H0** et on accepte H1.
2. si $\varepsilon_{calculée} < \varepsilon t_h$ alors **on accepte H0** et on rejette H1.

Si le / les échantillon(s) considérés sont **représentatifs** de la population, on pourra alors extrapoler le résultat obtenu à l'ensemble de la population.

7) Exemple :

Soient 2 populations : la population française et la population suédoise. On se demande si il y a la **même proportions d'individus ayant des yeux bleux en France et en Suède**. Pour cela, on constitue par TAS (tirage au sort) **2 échantillons A et B représentatifs** des populations françaises et suédoises. On obtient le tableau des données suivant :

Échantillon	A = français	B = Suédois	total
Yeux bleus	60	150	210
Yeux non bleus	140	50	190
total	200	200	400

1. définir H0 et H1 :
   - **H0** : il n'y a pas de différence observée entre les populations françaises et suédoises, c'est à dire qu'il y a la même proportions d'individus ayant des yeux bleux en France et en Suède.

   - **H1** : il existe une différence signitificative entre les populations françaises et suédoises, c'est à dire qu'il n'y a pas la même proportion de personnes ayant les yeux bleux en France et en Suède.

2. Déterminer le caractère des données à étudier/comparer :
   - nationalité française / suédoise → caractère **qualitatif**
   - couleur des yeux → caractère **qualitatif**

3. Choisir le test en fonction du type de données : Nous somme en présence de 2 caractères qualitatifs, on pourra donc choisir d'utiliser un **test de comparaison des pourcentages.**

4. Choisir le seuil d'erreur de 1ère espèce α : Ici, le risque de première espèce est fixé à **5%.**

5. Recueillir les données, calculer Z et utiliser la règle de rejet :
   - εt_h = 1,96 pour α = 5%
   - L'énoncé vous donne $\varepsilon_{calculée}$ = 10,09

6. Interpréter les résultats :
   - Au niveau des échantillons, $\varepsilon_{calculée} > \varepsilon t_h$ alors **on rejette H0** et **on accepte H1**. Ainsi, il n'y a pas la même proportion de personnes ayant les yeux bleux en France et en Suède. On constate que **les yeux bleus sont donc plus fréquents dans l'échantillon de Suédois**.

   - Les échantillons A et B étant **représentatifs**, on peut alors extrapoler à l'ensemble des deux populations et dire que **les yeux bleux sont plus fréquents dans la population Suédoise**.

II L'étude de la liaison entre deux caractères qualitatifs :

B Le test du χ^2 :

Soient **2 groupes A et B** et une caractéristique **qualitative x** (couleur des yeux etc.) . On peut se demande si la proportion d'individus du goupe A présentant x coïncide avec la proportion d'individus du goupe B présentant x.

1) définir H0 et H1.

- **H0** : il n'y a pas de différence observée entre les groupes A et B, c'est à dire que la la proportion d'individus du goupe A présentant x coïncide avec la proportion d'individus du goupe B présentant x.

- **H1** : il existe une différence signitificative entre les groupes A et B, c'est à dire que la la proportion d'individus du goupe A présentant x est différente de celle du goupe B.

2) Déterminer le caractère des données à étudier/comparer : Les variables sont :
- des types d'individus → variable **qualitative**
- une caractéristique x **qualitative**

3) Choisir le test en fonction du type de données En présence de données qualitatives, on peut choisir d'utiliser un **test du χ^2**.

4) Choisir le seuil d'erreur de 1ère espèce α généralement fixé à **5%.**

5) Recueillir les données, calculer Z et utiliser la règle de rejet
La variable Z est ici représentée par le **χ^2**. On comparera :
1. χ^2_{th} est donnée par la table du χ^2 en croisant :
   - α le risque de première espèce
   - **le nombre de degré de liberté.**
     Pour le test du χ^2 , le nombre de ddl vaut : **$(n_{lignes} - 1) \times (nb_{colonnes} - 1)$**
2. $\chi^2_{calculée} = \chi^2_{exp} = \sum \frac{(Oi - Ci)^2}{Ci}$

6) Interpréter les résultats : Au niveau de l'échantillon :
1. si $\chi^2_{calculée} \geq \chi^2_{th}$ alors **on rejette H0** et on accepte H1.
2. si $\chi^2_{calculée} \leq \chi^2_{th}$ alors **on accepte H0** et on rejette H1.

Si le / les échantillon(s) considérés sont **représentatifs** de la population, on pourra alors extrapoler le résultat obtenu à l'ensemble de la population.

7) Exemple :

Soient 2 populations : la population française et la population suédoise. On se demande si il y a la **même proportions d'individus ayant des yeux bleux en France et en Suède**. Pour cela, on constitue par TAS **2 échantillons A et B représentatifs** des populations françaises et suédoises. On obtient le tableau des données suivant :

Échantillon	A = français	B = Suédois	total
Yeux bleus	60	150	210
Yeux non bleus	140	50	190
total	200	200	400

1. définir H0 et H1 :
   - H0 : il n'y a pas de différence observée entre les populations françaises et suédoises, c'est à dire qu'il y a la même proportions d'individus ayant des yeux bleux en France et en Suède.
   - **H1** : il existe une différence signitificative entre les populations françaises et suédoises, c'est à dire qu'il n'y a pas la même proportion de personnes ayant les yeux bleux en France et en Suède.
2. Déterminer le caractère des données à étudier/comparer :
   - nationalité française / suédoise → caractère **qualitatif**
   - couleur des yeux → caractère **qualitatif**
3. Choisir le test en fonction du type de données : Nous somme en présence de 2 caractères qualitatifs, on pourra donc choisir d'utiliser un test du χ^2.
4. Choisir le seuil d'erreur de 1ère espèce α : Ici, le risque de première espèce est fixé à **5%.**
5. Recueillir les données, calculer de Z et utiliser la règle de rejet :
   - il y a (n_{lignes} – 1) x ($nb_{colonnes}$ – 1) = 1 x 1 = 1. Donc, pour α = 5%, $\chi^2_{Th} = 3{,}841$
   - L'énoncé vous donne $\chi^2_{calclué} = 81{,}2$
6. Interpréter les résultats :
   - Au niveau des échantillons, $\chi^2_{calclué} >>> \chi^2_{Th}$ alors **on rejette H0 et on accepte H1**. Ainsi, **il n'y a pas la même proportion de personnes ayant les yeux bleux en France et en Suède**. On constate que les yeux sont donc plus fréquents dans l'échantillon de Suédois.
   - Les échantillons A et B étant **représentatifs**, on peut alors extrapoler à l'ensemble des deux populations et dire que **les yeux bleux sont plus fréquents dans la population Suédoise**.

NB : Attention à toujours bien distinguer :
1. l'aspect *statistique* et ses conclusions
2. l'aspect *médical* et ses conclusions qui peuvent être différentes.

III L'étude de la liaison entre des caractères qualitatifs et quantitatifs :

Pour l'étude de la liaison entre un caractère qualitatif et un caractère quantitatif, on peut choisir d'utiliser soit :

3. un test de comparaison des moyennes
4. un test du T student

A Le test de comparaison des moyennes :

Soient **2 populations 1 et 2** et une caractéristique **quantitative** x (taille, poids etc.) de moyenne µ . On se demande si, en moyenne, la variable x des individus de la population 1 coincide avec celle des individus de la population 2. On met en place par TAS **2 groupes/échantillons représentatifs** des populations 1 et 2.

1) définir H0 et H1.

- **H0** : il n'y a pas de différence observée entre les groupes A et B, c'est à dire qu'en moyenne, la variable x des individus du groupe 1 coincide avec celle des individus du groupe 2.

- **H1** : il existe une différence signitificative entre les groupes A et B, c'est à dire qu'en moyenne, la variable x des individus du groupe 1 ne coincide pas avec celle des individus du groupe 2.

2) Déterminer le caractère des données à étudier/comparer : Les variables sont :

- des types d'individus → variable **qualitative**
- une caractéristique x **quantitative**

3) Choisir le test en fonction du type de données En présence de données qualitatives et quantitatives, pour n1 et n2 > 30, on peut choisir d'utiliser un test de **comparaison des moyennes**.

4) Choisir le seuil d'erreur de 1ère espèce α généralement fixé à **5%**.

5) Recueillir les données, calculer Z et utiliser la règle de rejet
La variable Z est ici représentée par **l'écart réduit ε**. On comparera :

1. εt_h donné par la table de l'écart réduit, en fonction de α
2. $\varepsilon_{calculée} = \varepsilon_{exp} = \dfrac{m_1 - m_2}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$

avec
m1 = moyenne du paramètre x calculée sur l'échantillon n°1
m2 = moyenne du paramètre x calculée sur l'échantillon n°2
s1 = écart type calculé sur l'échantillon n°1
s2 = écart type calculé sur l'échantillon n°2

6) Interpréter les résultats : Au niveau de l'échantillon :

1. si $\varepsilon_{calculée} > \varepsilon t_h$ alors **on rejette H0** et on accepte H1.
2. si $\varepsilon_{calculée} < \varepsilon t_h$ alors **on accepte H0** et on rejette H1.

Si le / les échantillon(s) considérés sont **représentatifs** de la population, on pourra alors extrapoler le résultat obtenu à l'ensemble de la population.

7) Exemple :

Soient 2 populations : la population des **hommes** et la population des **femmes**. On se demande si, *en moyenne, la taille des individus de la population des hommes coincide avec celle des individus de la population des femmes.* On met en place par TAS 2 groupes/échantillons représentatifs de 200 femmes et de 200 hommes. On obtient les moyennes suivantes :

Échantillon	Hommes	Femmes
n	n1=200	n2=200
m	m1 = 1,74 m	m2 = 1,68 m
s	S1 = 0,85	s2 = 0,8

1. définir H0 et H1 :
   - H0 : il n'y a pas de différence observée entre les populations des femmes et de hommes, c'est à dire qu'en moyenne, la taille des hommes coincide avec celle des femmes
   - H1 : il existe une différence signitificative entre les populations des femmes et des hommes, c'est à dire qu'en moyenne, la taille des hommes ne coincide pas avec celle des femmes. En moyenne les hommes sont plus grands que les femmes
2. Déterminer le caractère des données à étudier/comparer :
   - sexe → variable **qualitative**
   - taille → variable **quantitative**
3. Choisir le test en fonction du type de données : Nous sommes en présence de caractères **qualitatifs** et **quantitatifs**, avec $n1 = n2 > 30$, on pourra donc choisir d'utiliser un test de **comparaison des moyennes.**
4. Choisir le seuil d'erreur de 1ère espèce α : Ici, le risque de première espèce est fixé à **5%.**
5. Recueillir les données, calculer de Z et utiliser la règle de rejet :
   - $\varepsilon t_h = 1{,}96$ pour $\alpha = 5\%$
   - L'énoncé vous donne $\varepsilon_{calculée} = 8{,}81$
6. Interpréter les résultats :
   - Au niveau des échantillons, $\varepsilon_{calculée} > \varepsilon t_h$ alors **on rejette H0** et on accepte H1. Ainsi, en moyenne, **la taille des hommes de l'échantillon 1 ne coïncide pas avec celle des femmes de l'échantillon 2**. Les hommes de l'échantillon 1 sont plus grands que les femmes de l'échantillon 2.
   - Les échantillons 1 et 2 étant **représentatifs**, on peut alors extrapoler à l'ensemble des deux populations et dire que **d'une manière générale, les hommes sont plus grands que les femmes**.

III L'étude de la liaison entre des caractères qualitatifs et quantitatifs :

B Le test du T student :

Soient **2 populations 1 et 2** et une caractéristique **quantitative** x (taille, poids etc.) de moyenne μ . On se demande si, en moyenne, la variable x des individus de la population 1 coincide avec celle des individus de la population 2. On met en place par TAS **2 groupes/échantillons représentatifs** des populations 1 et 2.

1) définir H0 et H1.

- **H0** : il n'y a pas de différence observée entre les groupes A et B, c'est à dire qu'en moyenne, la variable x des individus du groupe 1 coincide avec celle des individus du groupe 2.

- **H1** : il existe une différence signitificative entre les groupes A et B, c'est à dire qu'en moyenne, la variable x des individus du groupe 1 ne coincide pas avec celle des individus du groupe 2.

2) Déterminer le caractère des données à étudier/comparer : Les variables sont :
- des types d'individus → variable **qualitative**
- une caractéristique x **quantitative**

3) Choisir le test en fonction du type de données En présence de données **qualitatives** et **quantitatives**, pour n1 ou n2 < 30, on peut choisir d'utiliser un **test de T student**

4) Choisir le seuil d'erreur de 1ère espèce α généralement fixé à **5%**.

5) Recueillir les données, calculer Z et utiliser la règle de rejet
La variable Z est ici représentée par l'écart réduit ε. On comparera :
1. le T student théorique donné par la table du T student. On l'obtient en croisant :
   - la valeur d'α
   - le nombre de ddl donné ici par (n1 – 1) + (n2 - 1)
2. T student calculé = $\dfrac{m_1 - m_2}{\sqrt{\dfrac{s^2}{n_1} + \dfrac{s^2}{n_2}}}$

avec
m1 = moyenne du paramètre x calculée sur l'échantillon n°1
m2 = moyenne du paramètre x calculée sur l'échantillon n°2
n1 = effectif de l'échantillon n°1
n2 = effectif de l'échantillon n°2
s = l'écart type commun aux 2 échantillons.

6) Interpréter les résultats :

Au niveau de l'échantillon :
1. si $T_{calculé} > T_{théorique}$ alors **on rejette H0** et on accepte H1.
2. si $T_{calculé} < T_{théorique}$ alors **on accepte H0** et on rejette H1.

Si le / les échantillon(s) considérés sont **représentatifs** de la population, on pourra alors extrapoler le résultat obtenu à l'ensemble de la population.

7) Exemple :

Soient 2 populations : la population des **hommes** et la population des **femmes**. On se demande si, *en moyenne, la taille des individus de la population des hommes coincide avec celle des individus de la population des femmes.* On met en place par TAS 2 groupes/échantillons représentatifs de 200 femmes et de 200 hommes. On obtient les moyennes suivantes :

Échantillon	**Hommes**	**Femmes**
n	N1=10 **(< 30)**	N2=10 (<30)
m	m1 = 1,74 m	m2 = 1,68 m
s	S = 0,8	s = 0,8

1. définir H0 et H1 :
   - **H0** : il n'y a pas de différence observée entre les populations des femmes et des hommes, c'est à dire qu'en moyenne, la taille des hommes coincide avec celle des femmes
   - **H1** : il existe une différence signitificative entre les populations des femmes et des hommes, c'est à dire qu'en moyenne, la taille des hommes ne coincide pas avec celle des femmes. En moyenne les hommes sont plus grands que les femmes
2. Déterminer le caractère des données à étudier/comparer :
   - sexe → variable **qualitative**
   - taille → variable **quantitative**
3. Choisir le test en fonction du type de données : Nous sommes en présence de caractères **qualitatifs** et **quantitatifs**, avec n1 < 30. On choisira donc d'utiliser un **test du T student**.
4. Choisir le seuil d'erreur de 1ère espèce α : Ici, le risque de première espèce est fixé à **5%.**
5. Recueillir les données, calculer Z et utiliser la règle de rejet :
   - α = 5 % et **nb de ddl = (n1-1) + (n2-1) = 9+9 = 18**. La table du T student, nous donne Tstudent th = 2,1
   - Tstudent calculé = 3,4 (*je précise que vous n'aurez jamais à le calculer, il sera donné dans l'énoncé*)
6. Interpréter les résultats :
   - Au niveau des échantillons, Tcalculé > Tth, alors **on rejette H0** et on accepte H1. Ainsi, en moyenne, **la taille des hommes de l'échantillon 1 ne coïncide pas avec celle des femmes de l'échantillon 2**. Les hommes de l'échantillon 1 sont plus grands que les femmes de l'échantillon 2.

Les échantillons 1 et 2 étant **représentatifs**, on peut alors extrapoler à l'ensemble des deux populations et dire que **d'une manière générale, les hommes sont plus grands que les femmes.**

III L'étude de la liaison entre des caractères qualitatifs et quantitatifs :

C Cas de séries appariées, méthode des couples :

On utilise la méthode des couples lorsqu'on étudie la liaison entre **deux variables qualitatives et quantitatives** dans 2 **échantillons non indépendants**.

Exemple : Soit un échantillon A de n patientes présentant une tumeur au sein. On s'intéresse à la *taille de cette tumeur avant et après un traîtement par chimiothérapie.*

- La taille de la tumeur est une variable **quantitative**
- Le traitement est une variable **qualitative**.

Les 2 échantillons A (avant traitement) et A' (après traitement) ne sont **pas indépendants**. Ils sont donc **appariées**. On utilisera alors la méthode des couples avec soit :

1. si n > 30, un test de **comparaison des moyennes** avec le paramètre écart réduit $\varepsilon = m_d / \sqrt{\frac{s2}{n}}$
2. si n < 30, un test de **Tstudent** avec le paramètre T : $t = m_d / \sqrt{\frac{s2}{n}}$

IV Etude de liaison entre des caractères quantitatifs :

Pour étudier la liaison entre 2 variables quantitatives, on utilise un coefficient de corrélation r. Soient une population et **2 caractéristiques quantatives x et y** (taille, poids, nombre de cigarettes etc.). On se demande s'il existe un lien entre x et y et si oui, lequel. On met en place par TAS un **échantillon représentatif** de la population.

1) définir H0 et H1.

- **H0** : il n'y a pas de lien entre les variables x et y.
- **H1** : il existe un lien entre x et y, celui-ci pouvant être :
    1. positif
    2. négatif

2) Déterminer le caractère des données à étudier/comparer :Les 2 variables sont des variables **quantitatives**.

3) Choisir le test en fonction du type de données En présence de données **quantitatives**, pour un nombre suffisant de sujets, on choisira d'utiliser un **coefficient de corrélation**.

4) Choisir le seuil d'erreur de 1ère espèce α généralement fixé à **5%.**

5) Recueillir les données, calculer Z et utiliser la règle de rejet

- On recueille différentes valeurs de x et de y.
- On trace la courbe **y = f(x)**
- On trace **la droite des moindres carrés**, c'est à dire la droite passant au plus près de chaque point de la courbe

- La **pente** de cette droite est appelé **coefficient de corrélation r**. Il est donné par la formule suivante :

$$r = \frac{\sum xy - \frac{\sum x . \sum y}{n}}{\sqrt{(\sum x^2 - \frac{(\sum x)^2}{n})(\sum y^2 - \frac{(\sum y)^2}{n})}}$$

Concernant r …

1. il est toujours compris dans **l'intervalle [-1;1]**
2. si il n'existe pas ou est nul, alors il n'y a **pas de corrélation** entre x et y au niveau de l'échantillon
3. si r existe et r > 0, alors il existe une **corrélation positive** entre x et y au niveau de l'échantillon
4. si r existe et r < 0, alors il existe une **corrélation négative** entre x et y au niveau de l'échantillon

6) Interpréter les résultats : Au niveau de l'échantillonsi |r calculé| > |r théorique| trouvé dans la table de r. alors, **on rejette H0**. Il existe bien un lien significatif entre x et y.

Si le / les échantillon(s) considérés sont **représentatifs** de la population, on pourra alors extrapoler le résultat obtenu à l'ensemble de la population.

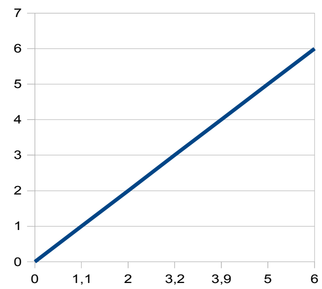
7) Exemple :

Soient la population des sujets **fumeurs** ET ayant **une tumeur au poumon**, et 2 caractéristiques **quantitatives** :

- x = nombre de cigarettes fumées par jour
- y = taille de la tumeur

On se demande s'il existe **un lien entre le nombre de cigarettes fumées et la taille de la tumeur et si oui, lequel**. On met en place par TAS, un échantillon **représentatif** de cette population. On relève pour chaque personne de l'échantillon les valeurs de x et y. A partir des données, on obtient :

- r théorique = 0,71
- le graphe suivant :



Parallèlement,

- r calculé = 0,78
- On a donc **r théorique < r calculé** → **on rejette H0**

Conclusion :

- Il existe bien une **corrélation entre le nombre de cigarettes fumées et la taille de la tumeur**
- Comme $r > 0$, cela signifie qu'il y a **corrélation positive**, c'est à dire que *plus le sujet a fumé de cigarettes, et plus la taille de la tumeur augmente.*
- Comme l'échantillon a été réalisé par **TAS**, on peut extrapoler les résultats à l'ensemble de la population des fumeurs présentant une tumeur pulmonaire.

V Les tests non paramétriques :

On utilisera des tests non paramétriques lorsque les **effectifs sont trop faibles** (inférieurs à **5**) pour utiliser les tests classiques et qu'au moins une des 2 variables est une variable quantitative. Dans ce cas, les populations ne se distribuent pas normalement et les tests non paramétriques présentent une **excellente robustesse**.

Ce tableau résume bien quel test utiliser en fonction de l'effectif :

Effectif	Données Quantitatives	Données Qualitatives	Données Qualitatives - Quantitatives
>4 & <12	r' de Spearman	Comp % ou χ^2	U Mann & Withney
>12 & < 30	Coeff de corrélation r	Comp % ou χ^2	t Student
> 30	Coeff de corrélation r	Comp % ou χ^2	Comp Moyennes

V Les tests non paramétriques :

A La liaison quantitatifs/ qualitatifs :test de U de Mann et Whitney

Soient 2 groupes de 5 individus chacun A et B et une caractéristique **quantitative** x (taille etc.) . On peut se demande si la proportion d'individus du goupe A présentant x coïncide avec la proportion d'individus du goupe B présentant x.

1) définir H0 et H1.

- **H0** : il n'y a pas de différence observée entre les groupes A et B, c'est à dire que la la proportion d'individus du goupe A présentant x coïncide avec la proportion d'individus du goupe B présentant x.

- **H1** : il existe une différence signitificative entre les groupes A et B, c'est à dire que la la proportion d'individus du goupe A présentant x est différente de celle du goupe B.

2) Déterminer le caractère des données à étudier/comparer : Les variables sont :

- des types d'individus → variable **qualitative**
- une caractéristique x **qualitative**

3) Choisir le test en fonction du type de données En présence de données **qualitatives** et **quantitatives**, et d'un effectif aussi réduit, on utilisera le **U de Mann et Whitney.**

4) Choisir le seuil d'erreur de 1ère espèce α généralement fixé à **5%.**

5) Recueillir les données, calculer Z et utiliser la règle de rejet
La variable Z est ici représentée par le **U de mann et Whitney** On comparera :

1. Uth donné par la table du U de Mann et Whitney en croisant :
   - nA
   - n_B - n_A, avec nB le plus grand des 2 effectifs

2. U calculé. (La méthode de calcul est exposée dans l'exemple qui suit)

6) Interpréter les résultats : Au niveau de l'échantillon :

1. si U calculé > U th, alors l'imbrication des 2 groupes est important et **on accepte H0**.
2. si U calculé < U th, alors l'imbrication des 2 groupes est très faible et **on rejette H0**

Si le / les échantillon(s) considérés sont **représentatifs** de la population, on pourra alors extrapoler le résultat obtenu à l'ensemble de la population.

7) Exemple :

Soient 2 populations : la population des **hommes** et la population des **femmes**. On se demande si, *en moyenne, la taille des individus de la population des hommes coincide avec celle des individus de la population des femmes.* . On met en place par TAS 2 groupes/échantillons représentatifs de 5 femmes et de 5 hommes. On obtient les moyennes suivantes :

Femmes (A)	1m58	1m60	1m65	1m66	1m68
Hommes (B)	1m67	1m69	1m75	1m80	1m90

1. définir H0 et H1 :
   - **H0** : il n'y a pas de différence observée entre les populations des femmes et de hommes, c'est à dire qu'en moyenne, la taille des hommes coincide avec celle des femmes
   - **H1** : il existe une différence signitificative entre les populations des femmes et des hommes, c'est à dire qu'en moyenne, la taille des hommes ne coincide pas avec celle des femmes. En moyenne les hommes sont plus grands que les femmes

2. Déterminer le caractère des données à étudier/comparer :
   - sexe → variable **qualitative**
   - taille → variable **quantitative**

3. Choisir le test en fonction du type de données : Nous sommes en présence de caractères **qualitatifs** et **quantitatifs**, avec $n_A = n_B = 5$. On choisira donc d'utiliser un **test de U Mann et Whitney**

4. Choisir le seuil d'erreur de 1ère espèce α : Ici, le risque de première espèce est fixé à **5%**.

5. Recueillir les données, calculer Z et utiliser la règle de rejet :
   - α = 5 % et nA-nB = 5 – 5 = 0. La table du U, nous donne U th = 2
   - Calcul de Ucalculé :

Théorie	Exemple
On classe toutes les valeurs **par ordre croissant** en fonction de leur appartenance à A ou à B.	**1,58 / 1,60 / 1,65 / 1,66 / 1,67/ 1,68/ 1,69 / 1,75/ 1,80 / 1,90**
On cherche le paramètres $\mathbf{U_{BA}}$. Pour chaque membre de A, on cumule le nombre de membre de B qui lui sont passés devant.	$\mathbf{U_{BA} = 0 + 0 + 0 + 0 + 1 = 1}$
NB : $U_{AB}+U_{BA} = n_A \times n_B$	**Ici,** $U_{AB}+U_{BA} = n_A \times n_B = 25$ Donc UAB = 25 – 1 = 24

6. Interpréter les résultats :
   - Au niveau des échantillons, Ucalculé < Uth, alors alors l'imbrication des 2 groupes est très faible et **on rejette H0.** Cela signifie qu'**il y a une différence de taille significative entre les hommes et les femmes**.

Les échantillons 1 et 2 étant **représentatifs**, on peut alors extrapoler à l'ensemble des deux populations et dire que **d'une manière générale**, **les hommes sont plus grands que les femmes**.

B Le coefficient r' de Spearman :

On l'utilise dans le cadre d'un **test de corrélation non paramétrique** . On utilise la même méthode que pour le coefficient de corrélation r, mais pour des échantillons de moins de 12 sujets et en utilisant la table théorique du r' de Spearman. La formule du coefficient r ' est :

$$r' = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$