Tutorat 2023-2024

FORMATION EN SOINS INFIRMIERS

PREFMS CHU DE TOULOUSE

Rédaction 2023-2024

UEC 20

Initiation à la recherche

Eléments statistiques de base

Rédigé par Sourd Dorian à partir du cours de J.SHOURIK présenté le 25/03/2024.

Eléments statistiques de base

I. Objectifs

Notion de statistiques descriptive :

Savoir produire et interpréter des statistiques descriptives élémentaires : effectifs, pourcentages, moyennes, écarts types, médianes et quantiles

Introduction à la théorie des tests statistiques :

Savoir interpréter les résultats d'un test bivarié de comparaison de pourcentages ou de moyennes (comprendre la nécessité de réaliser des tests d'hypothèse lorsque l'on travaille sur un échantillon, connaître les risques d'erreur de première et seconde espèce, comprendre la signification de la p-value, savoir interpréter un résultat statistiquement significatif et un résultat statistiquement non significatif)

II. Notion de statistiques descriptives

Pourquoi les statistiques en santé ?

- Définir des « normes » et distinguer le « normal » du « pathologique »
- Tester des associations, juger de la causalité
- Prendre en compte les phénomènes de confusion
- Mesurer/prévoir l'évolution d'un paramètre

a. Définition

Les unités statistiques sont les sujets faisant l'objet de l'étude.

Les variables statistiques sont des paramètres pouvant prendre différentes valeurs d'une unité statistique à l'autre. Les variables qualitatives sont des variables catégorielles. Il existe 2 types de variables qualitatives :

- Nominales : sans relation d'ordre (ex : rouge, bleu, jaune)
- Ordonnées : avec relation d'ordre (petit, moyen, grand)

Précision : une variable représente une valeur et ne peut être la combinaison de plusieurs valeurs. Par exemple, une variable ne peut pas représenter une petite pomme jaune. Il y a 2 variables : la variable petite et la variable jaune.

Il existe aussi des variables quantitatives qui peuvent être :

- Discontinue (=discrète) : ne prenant que des valeurs entières
- Continue : incluant une infinité de nombre

b. Représentations synthétiques d'une variable statistique

1) Variables qualitatives

Pour représenter une variable qualitative, on peut utiliser des tableaux de fréquence :

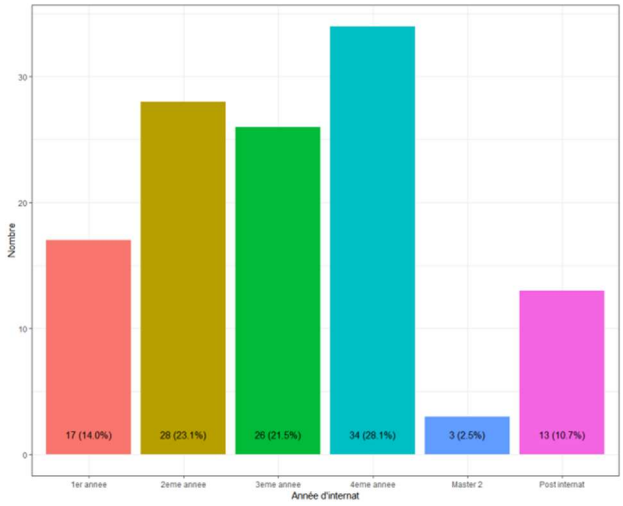
- La fréquence absolue : le nombre de cas
- La fréquence relative : pourcentage

Sur le tableau ci-contre, on a représenté les fréquences absolues par exemple le nombre de femme fumeuses : 70. On a aussi représenté des fréquences relatives par exemple le pourcentage d'homme fumeur : 53,3%.

Variable booléenne, dichotomique, binaire, à 2 modalités

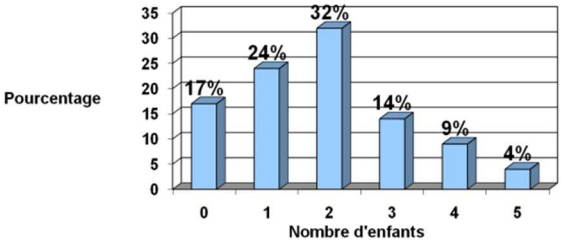
	N = 150
Sexe, n (%)	
hommes	80 (53,3 %)
femmes	70 (46,7 %)
Tabagisme, n (%)	
non fumeurs	77 (51,3 %)
anciens fumeurs	28 (18,7 %)
fumeurs	45 (30,0 %)

Une autre manière de représenter les données est le Barplot ou diagramme en barre :



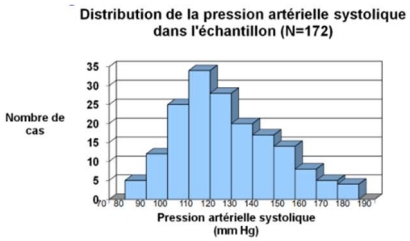
2) De la variable qualitative à la variable quantitative

Il existe maintenant des variables entre les deux. Le nombre d'enfant par exemple est une variable discrète et bornée.



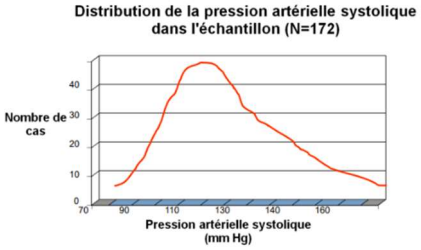
- Discrète : le nombre d'enfant ne peut avoir que des valeurs entières
- Bornée : il y a un nombre maximum d'enfant

On peut considérer cette variable comme quantitative : le nombre moyen d'enfant par exemple. Néanmoins, cette donnée n'est pas très intéressante. Ce qu'on peut voir c'est le nombre de personne qui ont 0 enfant, 1 enfant, 2 enfants, 3 enfants, 4 enfants et 5 enfants. Ce sont alors des variables discrètes puisqu'on ne peut pas avoir 1,3 enfant par exemple.



Un autre exemple est la distribution de la pression artérielle systolique dans un échantillon de 172 personnes dans l'histogramme ci-contre. La pression systolique est une variable qui peut prendre toutes les valeurs possibles dans son ensemble de définition. Par exemple : 125,3 mmHg. Ici on a des intervalles qui sont représentées.

On peut également synthétiser les valeurs par une courbe de distribution puisque cette variable peut prendre toutes les valeurs sur son espace de définition.



3) Variables quantitatives : indicateurs de tendance centrale

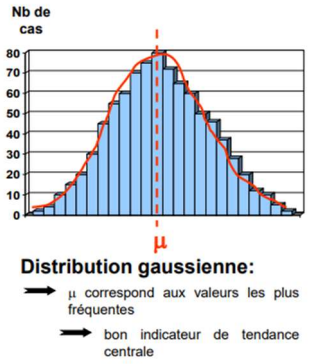
Pour les variables qualitatives, on peut les présenter sous la forme de fréquence et de pourcentage. Les variables quantitatives peuvent se présenter de différentes façons :

<u>On peut calculer une moyenne arithmétique</u>

La moyenne se calcul grâce à la somme des éléments divisée par leur nombre
Exemple : on a la liste de note suivante 13/20, 14/20, 14/20, 15/20 et 16/20.
La moyenne $\mu = \frac{13+14+14+15+1}{5}$ = 14,4.

$$\mu = \frac{1}{n} \sum_{i=1}^{n} X_i$$

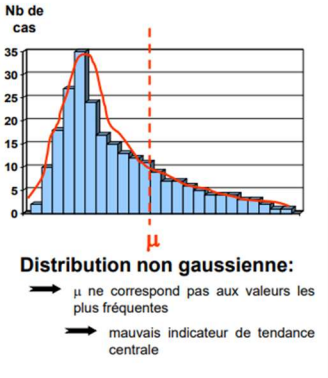
$$\mu = \frac{1}{n} (X_1 + X_2 + X_3 + \ldots + X_n$$



Une moyenne est intéressante lorsque les valeurs qui la composent sont assez bien centrées, lorsque la liste possède une distribution gaussienne.

La distribution gaussienne est symétrique sur le graphique. Sa moyenne est alors égale à son mode (modalité la plus représentée de la variable) : Le nombre de cas le plus élevé se situe au milieu, au niveau de la moyenne. Dans la distribution gaussienne, la moyenne est aussi à la médiane : autant de personne de chaque côté.

La moyenne μ correspond alors aux valeurs les plus fréquentes et c'est un bon indicateur de tendance centrale. C'est un bon indicateur de tendance centrale.



Dans une distribution non gaussienne, la moyenne ne correspond pas au mode, car le nombre de cas ne se situe pas au niveau de la moyenne, et elle ne correspond pas à la médiane car il n'y a pas de symétrie.

La moyenne μ ne correspond donc pas aux valeurs les plus fréquentes et c'est donc un mauvais indicateur de tendance centrale.

Les salaires sont un bon exemple de distribution non gaussiennes car un toute petite partie de la population gagne beaucoup d'argent et tirent le salaire moyen vers le haut.

La médiane

La médiane est une valeur centrale séparant l'échantillon en deux moitiés :

- 50% des valeurs sont au-dessus
- 50% des valeurs sont au-dessous
- Le rang de la médiane se trouve en calculant $(n + 1)/2$

Les médianes ne se divisent, ne se multiplient, ne s'additionnent et ne se soustraient pas. La différence de 2 moyennes est la moyenne des différences. La différence de 2 médianes n'est pas la médiane des différences.

Les informations en italiques sont un + pour la compréhension.

Le graphique en bleu est un diagramme en violon. Ils sont utilisés pour visualiser la distribution des données et leur densité de probabilités. C'est une combinaison entre un diagramme en boîte et un diagramme de densité.

Ce graphique associe en ordonnés une valeur de qualité de vie comprise entre 0 et 100 (« comment ça va aujourd'hui entre 0 et 100 ? ») et en abscisse la fréquence d'apparition.

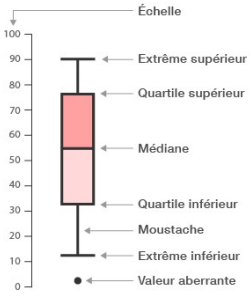


Diagramme en boîte

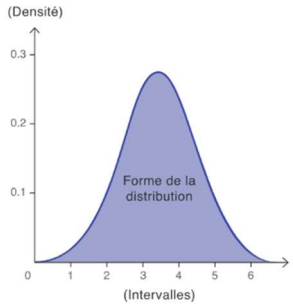


Diagramme de densité

On y retrouve les informations suivantes :

- La moyenne : représentée par le losange jaune au centre
- La médiane : représentée par la ligne noire entre les 2 carrés blancs
- La fréquence d'apparition : la largeur de l'aire bleue sur l'axe des abscisses (horizontal)
- Les intervalles de quartile : ici le 1er quartile par le carré blanc du haut et le 3e quartile par le carré blanc du bas.

Cette courbe n'est pas gaussienne car elle possède plusieurs modes locaux et il n'y a pas de courbe de distribution linéaire.

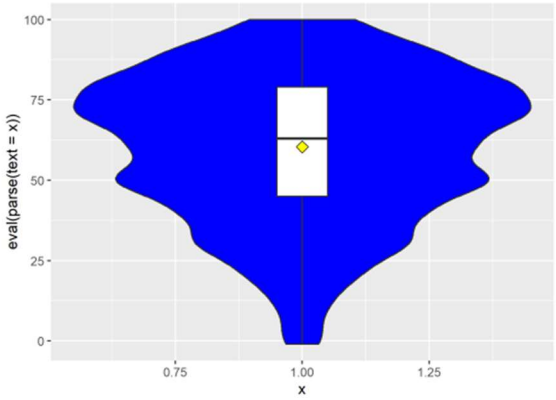


Diagramme en violon

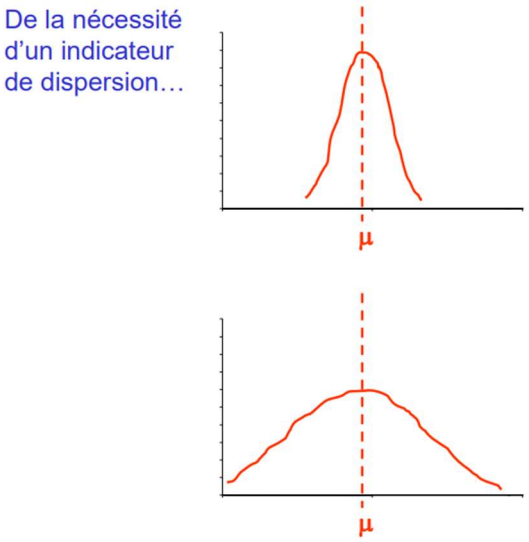
Exemples de représentation de la médiane :

Classement des sujets par ordre croissant de la valeur de la variable x

	Variable X		
sujet 1	15	sujet 4	10
sujet 2	12	sujet 16	11
sujet 3	13	sujet 2	12
sujet 4	10	sujet 6	12
sujet 5	17	sujet 15	12
sujet 6	12	sujet 3	13
sujet 7	15	sujet 19	13
sujet 8	15	sujet 9	14
sujet 9	14	sujet 1	15
sujet 10	18	sujet 7	15
sujet 11	20	sujet 8	15
sujet 12	15	sujet 12	15
sujet 13	17	sujet 17	15
sujet 14	16	sujet 14	16
sujet 15	12	sujet 20	16
sujet 16	11	sujet 5	17
sujet 17	15	sujet 13	17
sujet 18	18	sujet 10	18
sujet 19	13	sujet 18	18
sujet 20	16	sujet 11	20

Rang de la médiane :
(n+1) / 2 = (20 + 1) / 2 = 10.5

Valeur de la médiane : 15



Sur la liste à gauche, il y a un classement des sujets par ordre croissant de la valeur de la variable X comprise entre 10 et 20. Le rang de la médiane se calcul alors de la manière suivante : $(n+1)/2 \ = (20+1)/2 \ = 10{,}5$. On regarde ensuite au rang 10,5 (donc entre le 10e sujet et le 11e sujet et on obtient la valeur 15 pour la médiane. Cela signifie qu'il y a 50% des personnes qui ont au-dessus de 15 et 50% des sujets qui ont en dessous de 15. La moyenne de cette liste est différente en revanche : en calculant, on trouve une moyenne à 14,7.

Si on regarde les courbes de droite (on considère des tracés rigoureux), on remarque qu'elles ont les mêmes moyennes, elles ont une distribution gaussienne puisqu'elles ont le même mode et la même médiane. La différence entre ces 2 courbes est la dispersion : dans la courbe du haut, les valeurs sont rassemblées autour de la moyenne et dans la courbe du bas, les valeurs sont dispersées au-delà de la moyenne. Cette notion est importante pour connaître l'étendue possible d'une valeur. Si on prend la situation où on regarde la taille des étudiants d'un amphithéâtre associée à leur fréquence d'apparition. On considère que la moyenne des étudiants est de 1,6m. Si on dit que la majorité est comprise entre 1,55m et 1,65m, c'est en revanche différent si on dit que la majorité est comprise entre 1,20m et 2,10m.

4) Variables quantitatives : indicateurs de dispersion

La variance (Je ne vous demanderez pas de calculer des variances) :

Les indices de dispersion nous permettent de connaître l'écart de valeurs à la moyenne.

La variance

- La variance est la moyenne des carrés des écarts des valeurs par rapport à la moyenne.
- L'unité de la variance est l'unité de la variable étudiée au carré.

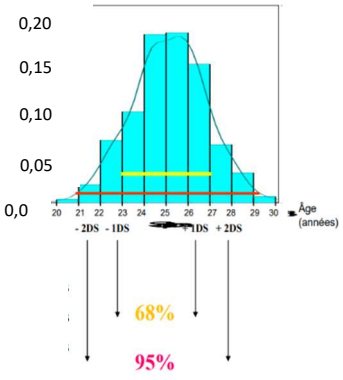
$$\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu)^2$$

L'écart type ou déviation standard (DS)

$$\sigma = \sqrt{\sigma^2}$$

L'écart est la racine carrée de la variance. Quand on a une distribution normale :

- 68% des observations sont comprises entre Moy μ +/- 1DS
- 95% des observations sont comprises entre Moy μ +/- 2DS
- 99% des observations sont comprises entre Moy μ +/- 3DS



Si on reprend l'exemple de l'étude de la taille des étudiants dans un amphithéâtre : on admet que la moyenne de la taille est de 1,7m et on considère que l'écart-type est de 5. On peut alors dire que 95% des personnes sont comprises entre 1,6m (1,7-2*5) et 1,8m (1,7+2*5).

Présenter un écart-type est intéressant si notre variable a une distribution normale.

Les quantiles

Les quantiles sont similaires à la médiane, mais au lieu d'être 50% au-dessus ou 50% en-dessous, on a des valeurs différentes. Si on prend une liste avec un effectif à 100. Pour le 75^{e} percentile, on a 75% en-dessous de la valeur et 25% au-dessus de la valeur du 75^{e} percentile. Autrement dit :

- (k – 1) valeurs séparant l'échantillon en k zones comportant le même nombre d'observations
  - k = 3 : Tertiles
  - k = 4 : Quartiles
  - k = 10 : Déciles
  - k = 100 : Centiles ou percentiles
- Le rang des quantiles est un multiple de (n + 1) / k
- Un intervalle entre deux quantiles correspond à un intervalle interquantile

sujet 4	10
sujet 16	11
sujet 2	12
sujet 6	12
sujet 15	12 ←
sujet 3	13
sujet 19	13
sujet 9	14
sujet 1	15
sujet 7	15 ←
sujet 8	15
sujet 12	15
sujet 17	15
sujet 14	16
sujet 20	16
sujet 5	17 ←
sujet 13	17
sujet 10	18
sujet 18	18
sujet 11	20

Si on étudie la liste ci-contre, on peut trouver les indications suivantes :

- Rand de la médiane : $(n+1)/2 \ = (20+1)/2 \ = 10{,}5$
- Valeur de la médiane : 15
- Rang des quartiles : $(n+1)/4 \ = (20+1)/4 \ = 5{,}25.$
  - 1er quartile : 5,25
  - 2^{e} quartile : 10,5
  - 3^{e} quartile : 15,75
- Valeurs des quartiles :
  - 1er quartile : 12,25 (flèche rouge du haut)
  - 2^{e} quartile : 15 (flèche rouge du milieu)
  - 3^{e} quartile : 16,75 (flèche rouge du bas)
- Intervalle interquartiles (noté IIQ) : [12,25 ; 16,75] (rappel : lorsque qu'on note 2 valeurs entre [], cela signifie qu'on parle d'une liste de nombre situés entre ces 2 valeurs.

Certaines données quantitatives sont des paramètres continus (ex le poids, la taille) On présentera une description précise de chacun des paramètres (la moyenne ou la médiane associé à un indice de dispersion (écart-type, valeurs extrême, espace interquartile).

Certaines données quantitatives sont des paramètres quantitatifs mais discrets (ex nombre de jours dans la semaine, nombre de traitements consommés dans une journée) On présentera une description précise de chacun des paramètres (n, %).

c. Données quantifiables

1) Binaires, ordinales et multinomiales

Binaire : Peut s'exprimer en présent ou absent (0 ou 1) : Décès, maladie

Ordinale ou multinomiale : Peut prendre plusieurs valeurs mutuellement exclusives : Petit/moyen/grand, rouge/bleu/vert

On présente l'effectif et la proportion de patient dans chaque catégorie. Exemple :

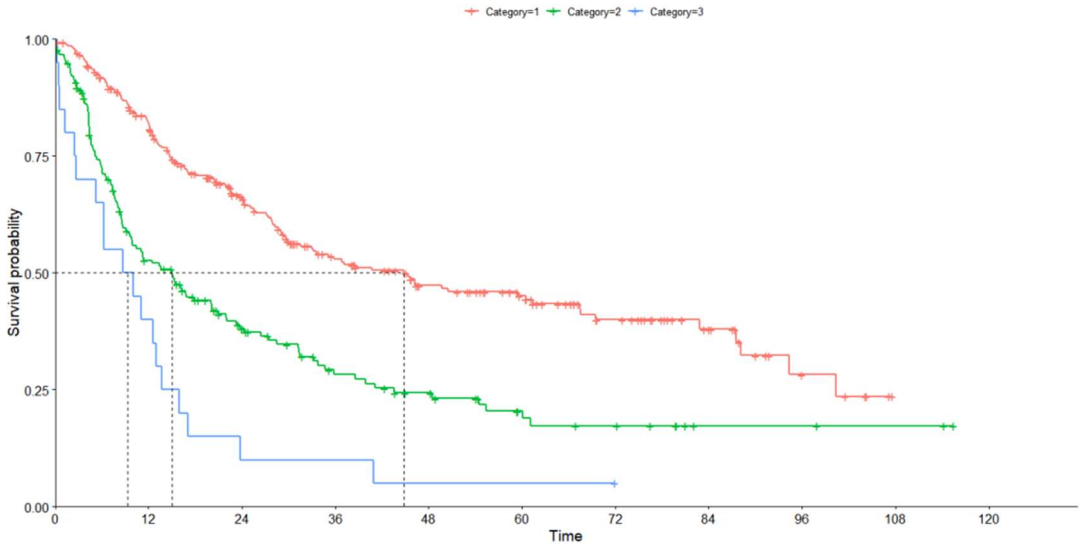
Variable	Nombre ou Moyenne	%	SD
Sex: F	432	76,3	
Sex: H	134	23,7	
Age	54		10,7
Lieu: Urbain	5	0,9	
Lieu: Rural	12	2,1	
Lieu: Semi-Urbain	160	28,3	
Teleconsultation_nombre: 1-5	66	15,5	
Teleconsultation_nombre: 5-10	132	30,9	
Teleconsultation_nombre: 11-20	92	21,5	
Teleconsultation_nombre: 21-50	97	22,7	
Teleconsultation_nombre: 50+	34	8	

On a ici représenté plusieurs variables :

- Le sexe : variable qualitative
- L'âge : variable quantitative (non discrète car on peut avoir 54,2 ans)
- La localisation : variable qualitative nominale
- Le nombre de téléconsultation : variable qualitative ordonné car on a mis des intervalles (1 à 5 consultations, 5 à 10 consultations…).

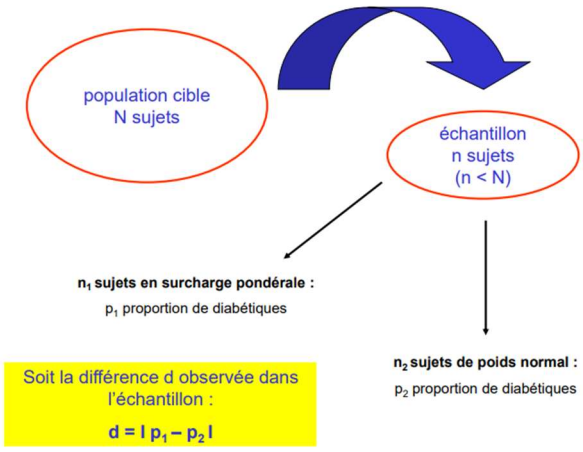
2) Prise en compte du temps

Le temps est le délai avant la survenue d'un événement. Il s'exprime sous forme de deux variables : une durée (quantitative) et la présence ou l'absence d'événement (binaire). Il se présente sous la forme de courbe de survie, de médiane de survie et de suivi.



Le graphique présenté ci-dessus est une courbe de survie. Au départ, tout le monde est vivant et au fur-et-à-mesure, les personnes décèdent. Les pointillés verticaux (correspondant à une valeur sur l'axe des abscisses) nous montrent le temps au bout duquel la moitié des personnes sont décédés pour chacune des 3 courbes. On appelle cela la médiane de survie.

III. Introduction à la théorie des test statistiques



Dans les tests statistiques, on cherche à savoir si les résultats obtenus sont représentatifs d'une population réelle.

Par exemple un objectif d'une étude peut être : Déterminer si la prévalence du diabète est différente chez les sujets en surcharge pondérale par rapport aux sujets de poids normal.

Pour répondre à cette problématique, on ne peut pas mesurer le poids de tous les diabétiques dans la population cible N. En revanche, on constituer un échantillon n d'enquête le plus représentatif possible et ensuite extrapoler les résultats à la population générale.

On recrute des sujets en surcharge pondérale et des sujets en poids normal. On peut se demander dans les sujets en surcharge pondérale, qu'elle est la proportion de diabétique. Cette proportion sera alors notée p_1. On peut aussi se demander dans les sujets de poids normal, quelle est la proportion de diabétique. Cette deuxième proportion sera alors notée p_2.

Lorsque l'on fait la différence de ces 2 proportions on a $D = |\, p1 - p2 \,|$. S'il n'y a pas de différence entre ces 2 proportions, on aura alors $d = 0$. Cela signifie qu'il y aura autant de personnes diabétiques chez les sujets en surcharge pondérale que chez les sujets de poids normal.

Cette hypothèse est nommée « hypothèse nulle H_0 » : La prévalence du diabète dans la population cible est identique parmi les sujets de poids normal et parmi les sujets en surcharge pondérale. *La probabilité associée à cette hypothèse est : « Quelle est la probabilité d'observer ce que j'observe s'il n'y a en réalité pas de différence*

$$D = |\, p1 - p2 \,|\; tend\ vers\ 0$$

On en décrit aussi une deuxième « l'hypothèse alternative H_1 » : La prévalence du diabète dans la population cible est différente parmi les sujets de poids normal et parmi les sujets en surcharge pondérale.

$$D = |\, p1 - p2 \,| \neq 0$$

(L'outil « | » nous permet de considérer le résultat uniquement en valeur absolue.

Si l'échantillon est de taille suffisante et représentatif :

- Sous H0 : $D = |p1 - p2|$ devrait être petite
- Sous H1 : $D = |p1 - p2|$ devrait être grande

Population cible / échantillon	absence de différence D → 0	existence d'une différence D ≠ 0	
d petite	Conclusion vraie Probabilité 1 - α	Conclusion fausse Risque β de seconde espèce	Conclure à tort à l'absence de différence entre les groupes
d grande	Conclusion fausse Risque α de première espèce	Conclusion vraie Puissance probabilité 1 - β	Probabilité de mettre en évidence une différence existant réellement entre les groupes
	Conclure à tort à une différence entre les groupes		

Risque α : risque de montrer une association qui n'existe pas (*C'est celui qui nous embête le plus*)

Risque β : risque de ne pas montrer une association qui existe. (*C'est le risque de ne pas conclure à une différence qui existe : Il y a en réalité une différence dans la population, mais on n'a pas réussi à la montrer »)*

La puissance se calcul en faisant 1 – β. C'est la probabilité de montrer une différence qui existe.

Le test statistique détermine si la différence d observée sur l'échantillon peut être considérée comme non due au hasard (différence existant réellement dans la population cible). Autrement dit, on réalise un test statistique pour savoir s'il est vraisemblable de rejeter l'hypothèse nulle H_0.

IV. Exemple

Une enquête épidémiologique a été mise en place pour étudier le lien entre tabagisme et cancer parmi des ouvriers d'une usine. N'ont été retenus que les dossiers pour lesquels on disposait d'un recul de 15 ans à compter du début de l'exposition. On a noté l'existence d'habitudes tabagiques et la survenue éventuelle d'un cancer du poumon dans les 15 ans. Sur 492 non-fumeurs, 2 cancers ont été observés, et sur 531 fumeurs, 17 cancers ont été observés.

Objectif de l'étude : Déterminer s'il existe un lien statistique entre tabagisme et cancer.

Sur ces données, pouvez-vous dire s'il existe un lien statistique entre tabagisme et cancer du poumon ?

- H0 : Absence de lien (fréquence du cancer non différente entre fumeurs et non-fumeurs), D = I P1 – P2 I proche de 0
- H1 : Présence d'un lien (fréquence du cancer différente entre fumeurs et non-fumeurs), D = I P1 – P2 I différent de 0

	Cancer +	Cancer -	
Tabac +	17	514	531
Tabac -	2	490	492
	19	1004	1023

- Fréquence du cancer parmi tabac + : 17 / 531 = 0,032 = 3,2 %
- Fréquence du cancer parmi tabac - : 2 / 492 = 0,004 = 0,4 %
- Test comparant ces deux pourcentages (test du Chi2, sous réserve des conditions d'application) :

p-value ≈ 0,001 p-value < 0,05 : on peut rejeter H0 et conclure avec un risque d'erreur de 0,001 (1 pour 1000) que la fréquence du cancer est différente chez les fumeurs comparativement aux non-fumeurs.

Autre exemple :

On souhaite comparer l'indice de masse corporelle moyen de sujets diabétiques et non diabétiques. On réalise pour cela une enquête épidémiologique sur un échantillon de 60 sujets, 30 diabétiques et 30 non diabétiques. Parmi les diabétiques, l'IMC moyen est de 27 kg/m2 (écart type 12), il est de 25 kg/m2 (écart type 11) parmi les non diabétiques.

Objectif de l'étude : Déterminer s'il existe un lien statistique entre IMC et diabète = comparer l'IMC moyen entre des sujets diabétiques et non diabétiques Sur ces données, pouvez-vous dire s'il existe un lien statistique entre IMC et diabète ?

- H0 : Absence de lien (moyenne d'IMC non différente entre diabétiques et non diabétiques), D = I M1 – M2 I proche de 0
- H1 : Présence d'un lien (moyenne d'IMC différente entre diabétiques et non diabétiques), D = I M1 – M2 I différent de 0

Moyenne d'IMC parmi les diabétiques : 27 kg/m2

Moyenne d'IMC parmi les non diabétiques : 25 kg/m2

Test comparant ces deux moyennes (test de Student sous réserve des conditions d'application) : p-value ≈ 0,30

p-value > 0,05 : on ne peut pas rejeter H0 (le risque d'erreur de première espèce est trop grand). On ne peut pas conclure sur cet échantillon à l'existence d'une différence d'IMC entre diabétiques et non diabétiques.

Deux explications possibles :

- Cette différence n'existe pas dans la réalité.
- Cette différence existe mais la puissance de l'étude était trop faible pour la mettre en évidence

V. Conclusion

La première chose à savoir est la variable qui nous intéresse. Les variables quantitative peut prendre n'importe qu'elle valeur.

Les variables quantitatives discrètes ne peuvent prendre que des nombres entiers

Les variables qualitatives ont des nombres réduits de modalités : malades/pas malade, homme/femme. Ces variables peuvent être ordonnée, on les nomme alors variables qualitatives ordinales, et elles peuvent être non-ordonnées, on les nomme alors variables multinomiales (ou nominales)

Les variables qualitatives n'ont rien à voir avec la recherche qualitative.

On peut transformer une variable quantitative en variable qualitative. Par exemple, la variable quantitative discrète qui représente le nombre d'enfant peut être catégorisée en 1 enfant, 2 enfants ou plus. On a alors transformé une variable quantitative discrète en une variable qualitative ordonnée. Pour décrire les variables qualitatives, on utilise les effectifs et les proportions.

Pour décrire une variable quantitative, on utilise un indicateur de centralité :

- La moyenne si la distribution est « normale »
- La médiane si la distribution n'est pas « normale »

On utilise aussi un indicateur de dispersion :

- L'intervalle interquartile qui représente ¼ des valeurs comprises en-dessous et ¾ des valeurs comprises au-dessus ou inversement (1/4 au-dessus et ¾ au-dessous)
- L'écart-type construit à partir de la variance (la moyenne des carrés des écarts à la moyenne). Il nous sert lorsque la distribution est normale. Dans ce cas-là, on sait que 95% de la population se situe entre -2 et +2 écart-type

Les tests statistiques sont là pour nous dire si la différence ou l'association qui est montrée dans la population d'échantillon est due au hasard ou si elle est bien présente dans la population cible.

On définit 2 types de risque :

- Le risque de montrer une différence qui n'existe pas
- Le risque de na pas montrer une différence qui existe

On représente généralement 1 – le risque : la puissance du test (probabilité de montrer une différence qui existe)

Lorsqu'on réalise un test statistique, on nous donne le p value, qui est la probabilité d'observer ce qu'on observe dans notre échantillon sous H_0, donc dans le cas où il n'y a pas vraiment de différence. Généralement, on prend comme risque α acceptable : 5%. On dit alors que si la p value est <5%, ça n'est alors probablement pas dû au hasard et il y a alors probablement une vraie différence dans la population. Par contre, si on ne montre pas de différence, on en peut pas conclure qu'il n'y a pas de différence. Cela veut peut-être dire que le test manque de puissance statistique. La puissance dépend du nombre de sujet, de la taille et de l'hétérogénéité de l'échantillon.