SECTION I

ITEM 16.1

Statistiques descriptives – les variables qualitatives

POURQUOI EST-CE IMPORTANT ?

Les statistiques descriptives sont toujours le point de départ de toutes les analyses. Il est donc essentiel de savoir les calculer ou les repérer et les interpréter quand elles sont données dans l'énoncé.

Elles sont très différentes suivant que la variable est qualitative ou quantitative et peuvent donc, quand elles sont données, aider à trouver le type de la variable. Dans cette fiche, seul le cas des variables qualitatives est abordé.

Les variables quantitatives sont traitées dans la fiche suivante.

RAPPELS DE COURS

Dans ce rappel, on considère une variable qualitative X à 3 modalités, A, B et C, observée chez n individus (taille de l'échantillon). Cela se généralise à un nombre quelconque de modalités.

TABLEAU DE CONTINGENCE

C'est la statistique descriptive fondamentale pour une variable qualitative. Pour chacune des modalités, il donne le nombre d'individus observés avec cette modalité. En général, il est présenté en colonnes : chaque colonne correspond à une modalité. Le total des valeurs est égal à la taille de l'échantillon, n (tableau 16.1.1).

On appelle n_A l'effectif de la modalité A. C'est forcément un entier positif compris entre 0 et n. Il en va de même pour n_B et n_C ; enfin, $n_A + n_B + n_C = n$.

Tableau 16.1.1

	X			
Modalité	A	B	C	*Total*
Effectif	n_A	n_B	n_C	n

TABLEAU DE FRÉQUENCES

Il est obtenu en divisant le tableau de contingence par la taille de l'échantillon. Il s'exprime en fractions (valeurs entre 0 et 1) ou, plus souvent, en pourcentages (valeurs entre 0 et 100 %) (tableau 16.1.2).

Chaque valeur dans une case correspond à la *fréquence* (f) de la modalité, aussi appelée la *proportion* (p) de la modalité.

On a $p_A = n_A/n$, $p_B = n_B/n$ et $p_C = n_C/n$; ce sont trois valeurs comprises entre 0 et 1 et $p_A + p_B + p_C = 1$.

Tableau 16.1.2

	X			
Modalité	A	B	C	*Total*
Fréquence	p_A	p_B	p_C	1 (soit 100 %)

À noter

Par abus de langage, on considère souvent synonymes « proportion » et « probabilité » ; attention cependant, en statistique descriptive on calcule une proportion à partir de l'échantillon observé. On s'en sert pour *estimer* une probabilité (voir Estimer une probabilité par une proportion, item 16.3).

À noter

Très souvent, l'énoncé donnera directement le tableau de contingence ou le tableau de fréquence, soit sous forme de tableau, soit sous forme textuelle.

Attention

Lorsque l'énoncé donne des proportions, il faut toujours les convertir en effectifs pour pouvoir faire les calculs ultérieurs. Dans ce cas, du fait des arrondis, il peut y avoir quelques effectifs non-entiers : arrondissez-les à l'entier le plus proche, en vérifiant que le total vaut bien la valeur indiquée.

Pour aller plus loin

On peut définir, pour chaque modalité M observée avec la fréquence p_M, la cote (ou, en anglais, *odds*) associée, $c_M = p_M/(1 - p_M)$. Quoique rarement utilisée en tant que telle (sauf dans le milieu sportif), c'est une notion fondamentale en épidémiologie et en santé publique, où le rapport des cotes joue un rôle clef (voir Transversale 7, Rapport des cotes (ou *odds ratio*), item 5.5).

EXEMPLES

À partir des extraits d'exercices de concours ci-dessous, reconstruisez le tableau de contingence et le tableau des fréquences.

CONCOURS DÉCEMBRE 2014

EXERCICE 5

On se propose d'évaluer l'influence de l'alcool sur les accidents vasculaires cérébraux (AVC). Pour cela, 60 000 femmes âgées de 35 à 59 ans sont suivies régulièrement pendant 5 ans. Leur consommation moyenne quotidienne d'alcool est évaluée et convertie en quantité d'alcool pur ingérée (en g). Les résultats sont les suivants :

▶ 75 femmes ont fait un AVC parmi les 30 000 qui ne boivent pas d'alcool ;

▶ sur les 10 000 femmes ingérant plus de 15 g d'alcool pur par jour, 450 ont fait un AVC ;
▶ parmi les 20 000 femmes ingérant moins de 15 g d'alcool pur par jour, 210 ont fait un AVC.
Construisez le tableau de contingence correspondant à chacune des deux variables de cette étude.

CORRIGÉ

Il y a ici deux variables qualitatives : le fait de faire, ou non, un AVC (variable qualitative binaire) et la consommation d'alcool, classée en trois groupes (aucune, moins de 15 g/j, plus de 15 g/j ; variable ordinale ternaire). L'effectif total vaut $n = 60\,000$.
Pour la consommation d'alcool, les effectifs de chacune des trois modalités sont donnés à la lecture du texte et le tableau de contingence s'exprime

	Consommation			
Modalité	0	< 15 g/j	> 15 g/j	*Total*
Effectif	30 000	20 000	10 000	60 000
Fréquence	0,50	0,33	0,17	1

La dernière ligne reprend les fréquences. La variable étant ordinale, on prendra soin de bien respecter l'ordre des modalités dans le tableau.
Pour la survenue ou non d'un AVC, les valeurs figurent aussi dans le texte, mais réparties pour chaque groupe de consommation. Le nombre total d'AVC est donc 75 + 450 + 210 = 735 ; par différence, il y a donc 60 000 – 735 = 59 265 femmes qui n'ont pas d'AVC et les tableaux de contingence et de fréquences sont donc :

	Survenue d'un AVC		
Modalité	Non	Oui	*Total*
Effectif	59 265	735	60 000
Fréquence (%)	98,8	1,2	100

CONCOURS 2008-2009 « SUD »

EXERCICE 2

Dans le cadre d'une enquête sur le suivi des malades traités pour asthme, une étude a été réalisée sur un échantillon aléatoire de 32 sujets asthmatiques. Les variables retenues pour cette étude sont :
▶ le sexe (F ou M) et l'âge (ans) du patient ;
▶ la gravité clinique : asthme intermittent ou persistant léger (A1) ; asthme persistant modéré ou sévère (A2).
Les résultats avant traitement sont donnés dans le tableau ci-dessous.

Patient	**12**	**14**	**17**	**19**	**27**	**31**	**32**	**1**
Sexe	F	F	F	F	F	F	F	M
Gravité	A1	A1	A1	A1	A1	A1	A1	A1
Patient	**4**	**10**	**15**	**18**	**21**	**22**	**24**	**25**
Sexe	M	M	M	M	M	M	M	M
Gravité	A1	A1	A1	A1	A1	A1	A1	A1
Patient	**30**	**2**	**5**	**6**	**7**	**9**	**20**	**23**
Sexe	M	F	F	F	F	F	F	F
Gravité	A1	A2	A2	A2	A2	A2	A2	A2
Patient	**26**	**28**	**3**	**8**	**11**	**13**	**16**	**29**
Sexe	F	F	M	M	M	M	M	M
Gravité	A2	A2	A2	A2	A2	A2	A2	A2

Construisez le tableau de contingence du caractère homme/femme ; de la gravité clinique.

CORRIGÉ

Le sexe (« caractère homme/femme ») est une variable qualitative binaire. Les données étant indiquées patient par patient, il convient de compter le nombre d'hommes parmi ces 32. On notera que les patients sont triés, par gravité d'abord puis par sexe pour une gravité donnée : cela facilite le décompte… On dénombre ainsi 7 + 9 = 16 femmes et 10 + 6 = 16 hommes, la somme faisant bien 32 : le tableau de contingence de la variable « sexe » est donc :

	Sexe		
Modalité	Homme	Femme	*Total*
Effectif	16	16	32

De la même manière, on dénombre 17 patients avec une gravité A1 et 15 avec une gravité A2, conduisant au tableau de contingence suivant :

	Gravité		
Modalité	A1	A2	*Total*
Effectif	17	15	32

CONCOURS 2019

EXERCICE 3

Un essai thérapeutique randomisé a été réalisé pour comparer la qualité de vie de deux groupes de 100 patients chacun, souffrant d'insuffisance cardiaque chronique. Le groupe A bénéficiait de la prise en charge clinique habituelle en ambulatoire tandis que le groupe B bénéficiait d'un programme d'éducation thérapeutique.

27 % des patients du groupe A et 15 % des patients du groupe B vivaient seuls. Construisez les tableaux de concordance et de fréquences du fait de vivre seul.

CORRIGÉS

La variable « vivre seul » est une variable qualitative binaire. Dans le groupe A, comptant 100 patients, il y a 27 % des patients, soit 0,27 × 100 = 27 patients, vivant seuls ; dans le groupe B, 0,15 × 100 = 15 patients vivant seuls – soit un total de 27 + 15 = 42 patients vivant seuls ; le tableau de contingence est donc celui ci-dessous et l'on en déduit le tableau de fréquences (dernière ligne du tableau ci-dessous).

	Vivre seul		
Modalité	Non	Oui	*Total*
Effectif	158	42	200
Fréquence (%)	79	11	100