Le but de cette fiche est de présenter diverses situations dans lesquelles on peut utiliser un test du Khi2. L'intérêt de ce test est qu'il est assez 'passe-partout' car il requiert peu d'hypothèses (par exemple pas d'hypothèses de normalité.) En contrepartie, il sera moins précis que d'autres tests (comme par exemple les tests de comparaison de moyenne vus dans les fiches précédentes).

Dans tous les cas, la mise en oeuvre du test est similaire à celle déjà rencontrée dans les fiches précédentes :

- On formule une hypothèse nulle H_0
- On calcule à l'aide de nos observations la valeur $d^2 = \sum \frac{(O_i - T_i)^2}{T_i}$.
  (les O_i sont les effectifs observés et les T_i les effectifs théoriques sous H_0.
- On sait que sous l'hypothèse H_0, d^2 est une réalisation d'une variable aléatoire D^2 suivant une loi du Khi2 ayant un nombre ν de degrés de libertés.
  On détermine donc à l'aide d'une table de la loi du Khi-deux la valeur du seuil $Khi2_{seuil}$ pour le risque de 5%.
- On conclut :
  Si $d^2 > Khi2_{seuil}$, on rejette H_0 (au risque de 5% de le rejeter à tord) ; sinon, on accepte H_0 en indiquant 'les données observées ne permettent pas de rejeter H_0' (ce qui ne veut pas dire que H_0 est vraie...mais on fait comme si elle l'était)

(Noter que, pour que le test soit valide, il est nécessaire que tous les T_i soient supérieurs ou égaux à 5.))

1 Test d'indépendance

(D'après sujet biologie ens 2009)
On souhaite étudier l'influence du transgène HSP70-Alk3-γCrys-GFP sur la régénération de la queue d'un groupe de tétards, qui ont été amputés de 50% de leur queue.

On observe les résultats suivants au sujet de la régénération :

	Aucune	Partielle	Totale
Avec transgène	3	18	11
Sans transgène	46	14	33

(Ceci est le tableau contenant les effectif observés O_i.)

Question : Peut-on conclure, au seuil de risque de 5%, que le transgène a une influence significative sur la repousse ?

Réponse : Ici H_0 est 'la repousse de la queue et le traitement par le transgène sont indépendantes'.

On a :

	Aucune	Partielle	Totale		total
Avec transgène	3	18	11		32
Sans transgène	46	14	33		93
total	49	32	44		125

Sous l'hypothèse d'indépendance H_0, on aurait donc les effectifs théoriques :

Aucune	Partielle	Totale	total
Avec transgène	$\frac{32\times49}{125}$	$\frac{32\times32}{125}$	$\frac{32\times44}{125}$
Sans transgène	$\frac{93\times49}{125}$	$\frac{93\times32}{125}$	$\frac{93\times44}{125}$

(Ceci est le tableau contenant les effectif théoriques T_i.)

On vérifie que tous les effectifs théoriques sont bien supérieurs ou égaux à 5.
En effectuant une somme sur les 6 cases des tableaux on obtient $d^2 = 25.16$.

Or, dans le cas présent, le nombre de degrés de liberté est 2 ($(c-1)(l-1) = (3-1)(2-1)$), c désignant le nombre de lignes, l désignant le nombre de colonnes du tableau.) et on lit dans la table du khi2 que la valeur du seuil pour le risque de 5% est 5.991. d^2 étant supérieur à cette valeur, on rejette H_0 au risque de 5%.

2 Test d'adéquation à une loi donnée

NB : Dans le cas de lois continues, il vaut mieux utiliser la fonction de répartition empirique que le test du Khi2 qui nécessite un regroupement en classes.

2.1 La loi est entièrement déterminée à l'avance

On dispose d'un dé et on souhaite déterminer s'il est ou non truqué.
On le lance 50 fois et on note les résultats obtenus .

Numéro obtenu	1	2	3	4	5	6
Nombre de fois	4	9	10	7	9	11

(Ceci est le tableau contenant les effectif observés O_i.)

L'hypothèse que le dé est équilibré peut se traduire par H_0 :'la variable aléatoire égale au numéro obtenu lorsqu'on lance une fois le dé suit une loi uniforme sur $[\![1, 6]\!]$'.

Sous H_0, on aurait donc les effectifs théoriques :

Numéro obtenu	1	2	3	4	5	6
Nombre de fois	$\frac{50}{6}$	$\frac{50}{6}$	$\frac{50}{6}$	$\frac{50}{6}$	$\frac{50}{6}$	$\frac{50}{6}$

(Ceci est le tableau contenant les effectif théoriques T_i.)

On vérifie que tous les effectifs théoriques sont bien supérieurs ou égaux à 5.
En effectuant une somme sur les 6 cases des tableaux on obtient $d^2 = 3.76$.
Or, dans le cas présent, le nombre de degrés de liberté est 5 = 6-1 (tableau uniligne de 6 cases) et on lit dans la table du khi2 que la valeur du seuil pour le risque de 5% est 11.07.
d^2 étant inférieur à cette valeur, au risque de 5%, on ne rejette pas H_0 .

2.2 Certains paramètres de la loi ont dû être estimés

La différence avec le cas précédent est le nombre de degrés de liberté, qui diminue d'un nombre égal au nombre de paramètres estimés.
On a relevé le nombre de pieds de vulpin présents dans 100 carrés d'un mètre de côté pris au hasard dans une parcelle.

x_i	0	1	2	3	4	5	6	7 et plus
O_i	12	28	19	13	12	8	6	2

Question : La variable aléatoire égale au nombre de pieds de vulpin dans un tel carré suit-elle une loi de Poisson ?

Démarche à suivre :
Pour pouvoir calculer nos effectifs théoriques, il est nécessaire d'avoir le paramètre de cette loi de Poisson. Comme le paramètre d'une loi de Poisson est égal à son espérance, on l'estime par la moyenne de notre échantillon $\bar{X} = 2.43$.
Or , pour une loi de Poisson de paramètre 2.43 on a les probabilités suivantes :

x_i	0	1	2	3	4	5	6	7et plus
p_i	0.08804	0.21393	0.26	0.21054	0.1279	0.06216	0.02518	0.0123

d'où le tableau des effectifs théoriques :

x_i	0	1	2	3	4	5	6	7et plus
T_i	8.804	21.393	26	21.054	12.79	6.216	2.518	1.23

Or on a ici des effectifs théoriques inférieurs à 5 dans certaines classes ; il faut faire un choix de regroupement. Je regroupe les trois dernières et on a donc maintenant :

x_i	0	1	2	3	4	5 et plus
O_i	12	28	19	13	12	16
T_i	8.804	21.393	26	21.054	12.79	9,9668

La valeur de d^2 est ici 11.86.
Le nombre de degrés de liberté est de 4 = 6-1-1 (6 classes et un paramètre estimé).
Pour un risque de 5%, le seuil lu dans la table est 9.488 .
Comme la valeur observée est supérieure à ce seuil, au risque de 5% on rejette l'hypothèse que le nombre de pieds de vulpins dans un carré de 1m de côté de cette parcelle suit une loi de Poisson.
(Ce qui au passage permet de rejeter l'hypothèse que les localisations des pieds de vulpin sont inédépendantes les unes des autres ; voir cours de probas....)

3 Remarques diverses

1. Si les effectifs théoriques de certaines classes sont inférieurs à 5, il faut regrouper des classes entre elles. (Il y a cependant perte d'information.)
2. Le test du Khi2 peut nous dépanner dans le cas où l'on voudrait comparer 2 moyennes mais où les hypothèses de normalité ne sont pas satisfaites.

   (Par exemple on veut savoir si la taille d'une plante dépend du fait qu'elle soit cultivée avec ou sans une certaine substance.)
   L'idée est de regrouper les valeurs en classes de manière à obtenir un tableau du style :

Taille en cm	$[1,3[$	$[3,4[$	$[4,5[$	$[5,7[$	plus de 7
Avec substance	4	9	10	7	9
Sans substance	12	6	11	4	1

   On applique ensuite le test du Khi2 d'indépendance à ces valeurs. (Ici , le nombre de ddl est (2-1)*(5-1) = 4)

   Remarque : On peut aussi vouloir savoir si la taille d'une plante dépend non seulement du fait qu'elle soit cultivée avec ou sans une certaine substance , mais aussi de la concentration dans le sol de cette substance , ce qui mènerait à un tableau du type :

Taille en cm	$[1,3[$	$[3,4[$	$[4,5[$	$[5,7[$	plus de 7
Sans substance	4	9	10	7	9
Substance avec concentration faible	12	6	11	4	1
Substance avec concentration moyenne	13	6	1	4	1
Substance avec concentration forte	12	6	11	4	1

   On applique ensuite le test du Khi2 d'indépendance à ces valeurs.
   (Ici , le nombre de ddl est (5-1)*(4-1) = 12)
3. Le test du Khi2 n'est applicable que lorsque l'effectif théorique de chaque classe est supérieur ou égal à 5, ce qui amène parfois à devoir regrouper des classes.
   Si ce regroupement est gênant (par exemple plus assez de classes), on peut utiliser une autre version du test du Khi2, mais à réserver aux 'experts' :

Si moins de 20% des classes ont des effectifs théoriques compris entre 2 et 5 (exclus) on peut appliquer le test du Khi2 corrigé de Yates, où $d^2 = \sum \frac{(|O_i - T_i| - 0.5)^2}{T_i}$..