Prélude

Retour sur l'article de François Héran

- Dans l'enquête de François Héran, les chefs de ménage hommes possèdent en moyenne 0,56 chiens. Les chefs de ménages de femmes possèdent 0,27 chiens.
  - Est-ce différent ?
  - Est-on prêt à croire que, dans la population globale, il existe une différence du nombre de chiens possédés en fonction du sexe du chef de ménage ?
- Les hommes possèdent 0,39 chats contre 0,25 chats pour les femmes.
  - Femmes et hommes se différencient-ils plus par le nombre de chiens possédés que par le nombre de chats ?
- 53% des sans diplôme possèdent un animal domestique contre 44% des diplômés supérieurs au bac.
  - Que peut-on dire de cette différence ?

La solution statistique

- Connaître la dispersion autour de la moyenne
- Faire un test
- Le test : une manière (imparfaite) de mesurer le risque que l'on prend lorsque l'on fait confiance dans une différence statistique apparente

Introduction aux tests statistiques

Olivier Godechot

Qu'est-ce qu'un test ?

- Tester une proposition en statistique, c'est :

« Montrer que les contradicteurs ont peu de chance d'avoir raison »

ou bien

« Calculer la probabilité que le contraire de ce que l'on pense soit vrai ! »

- Et montrer que cette probabilité est très petite.
- Une espèce de raisonnement par l'absurde : un raisonnement par double négation

Un test de student : le nombre moyen de chiens possédés diffère-t-il en fonction du sexe du chef de ménage?

Statistiques							
Variable	**CSEX**	**Nb**	**Moyenne**	**Écart-type**	**Erreur Std**	**Minimum**	**Maximum**
chienn	**1**	4579	0,5628	0,8846	0,0131	0	11
chienn	**2**	1303	0,2678	0,6292	0,0174	0	6
chienn	**Diff (1-2)**		0,2949	0,8349	0,0262		

Tests de Student					
Variable	**Méthode**	**Variances**	**DF**	**Valeur du t est t**	**Pr > |t|**
chienn	**Pooled**	Equal	5880	11,25	<,0001
chienn	**Satterthwaite**	Unequal	2916	13,54	<,0001

- La probabilité de se tromper en rejetant l'égalité du nombre moyen de chiens possédés quel que soit le sexe du chef de ménage, est inférieure à 1 pour 10 000

Un test du chi-deux. La possession de chien(s) dépend-elle de l'habitat ?

Fréquence Pourcentage Pourct. en ligne Pourct. en col.	chien	pas de chien	Total
Immeuble	458	2157	2615
	7.80	36.74	44.54
	17.51	82.49	
	22.16	56.70	
Maison	1609	1647	3256
	27.41	28.05	55.46
	49.42	50.58	
	77.84	43.30	
Total	2067	3804	5871
	35.21	64.79	100.00

Fréquence manquante = 11

Statistique	DF	Valeur	Proba.
Khi-2	1	647.0328	<.0001

- Lecture littérale du test : *On a moins d'une chance sur 10000 de se tromper en rejetant l'idée que la possession ou non de chien(s) ne dépend pas du type d'habitat.*
- Lecture moins littérale : La situation est significativement différente (au seuil 1/10000) de l'indépendance.
- Ou encore : Le taux de possession de chien est significativement supérieur dans une maison que dans un immeuble

Pourquoi tester ?

- En statistiques, en général on travaille sur des échantillons et non sur la population complète.
- Même quand on travaille sur la population complète, on considère parfois que la population complète n'est qu'une réalisation de l'univers infini des possibles.
- Par conséquent les grandeurs observées sur l'échantillon (moyenne, écart-type, etc.) représentent plus ou moins bien ces mêmes grandeurs à l'échelle théorique, i.e. à l'échelle de la population [ou de l'univers des possibles].
- On fait alors des hypothèses sur ces grandeurs au niveau théorique et on mesure la compatibilité des réalisations empiriques avec les hypothèses théoriques.

Le contraire de ce que l'on pense

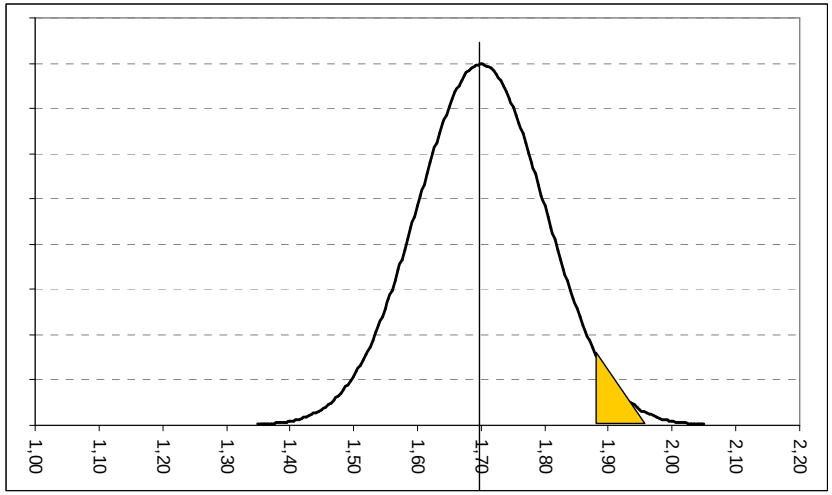
- D'abord on établit une hypothèse sur le paramètre au niveau théorique : On l'appelle H1
  - En général une différence (**≠**) ou une inégalité (< ou >)
    - Exemple1. H1: $m1_{théo} \neq m2_{théo}$
    - Exemple2. H1: $m1_{théo} > m2_{théo}$
- Le contraire de ce que l'on pense : c'est H0, l'hypothèse nulle, celle que l'on veut "nullifier" [to nullify] !
  - C'est une égalité (**=**) ou une inégalité (< ou >)
    - Exemple1. H0: $m1_{théo} = m2_{théo}$
    - Exemple2. H0: $m_{théo}1 < m2_{théo}$

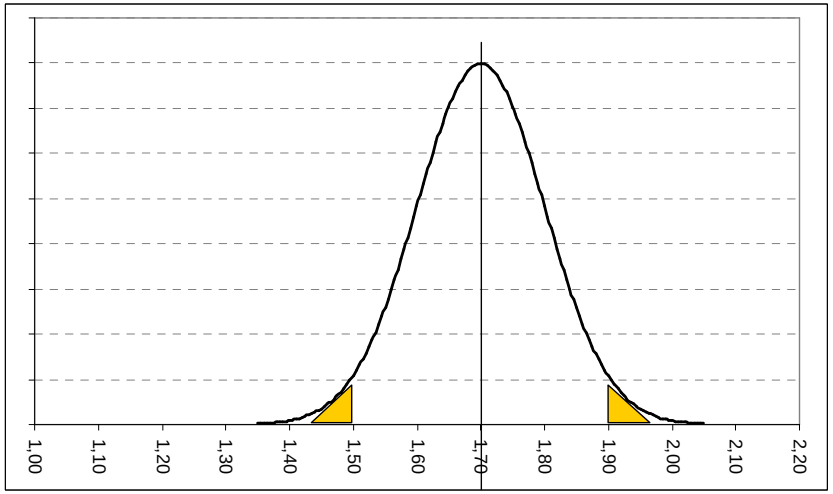
La probabilité que le contraire soit vrai

- Il faut faire des hypothèses sur la valeur des paramètres théoriques dans le cadre de H0. Par exemple, la moyenne théorique est nulle (et le cas échéant l'écart-type théorique est de tant…)
- On se dote d'une loi de probabilité ou d'une table de probabilité mesurant la répartition des écarts autour des paramètres de H0.
- Ensuite on calcule, sous les hypothèses de H0, la probabilité lorsque l'on génère des échantillons (de même taille que celui de notre échantillon théorique) d'obtenir des écarts à l'hypothèse nulle au moins aussi grands que l'écart entre les données empiriques et les paramètres de H0.
- Si cette probabilité est faible (<10%). On dit qu' « on rejette H0 au seuil de x% ». On a alors x% de chance de se tromper en acceptant H1.
- On dit alors généralement que la différence est significative.
- Si cette probabilité est importante (>10%), « on ne peut pas rejeter H0 ». Peut-on alors accepter H0 ? C'est une question discutée.... (Fisher vs Neyman Pearson).

Test unilatéral – test bilatéral

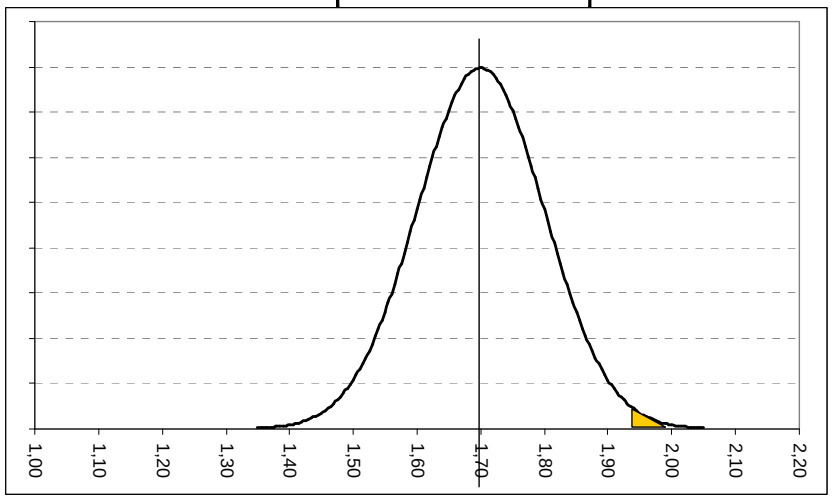
- Supposons pour H0 que notre paramètre $m_{théorique}$ suit une loi normale de moyenne 1,70 et d'écart-type 0,1 (par exemple la taille en m)
- A gauche on pose H0: $m_{théorique} < 1{,}70$ et à droite H0: $m_{théorique} = 1{,}70$.
- Les valeurs empiriques permettant de rejeter H0 au seuil de 5% sont marquées par les zones en jaune.
- A gauche, une moyenne empirique sur un échantillon supérieure à 1,86 permet de rejeter H0, en revanche des valeurs très faibles ne le permettent pas.
- A droite on rejettera H0 pour une moyenne empirique soit supérieure à 1,90 soit inférieure 1,50. D'un côté donné de la courbe, on n'est plus exigeant pour rejeter H0.

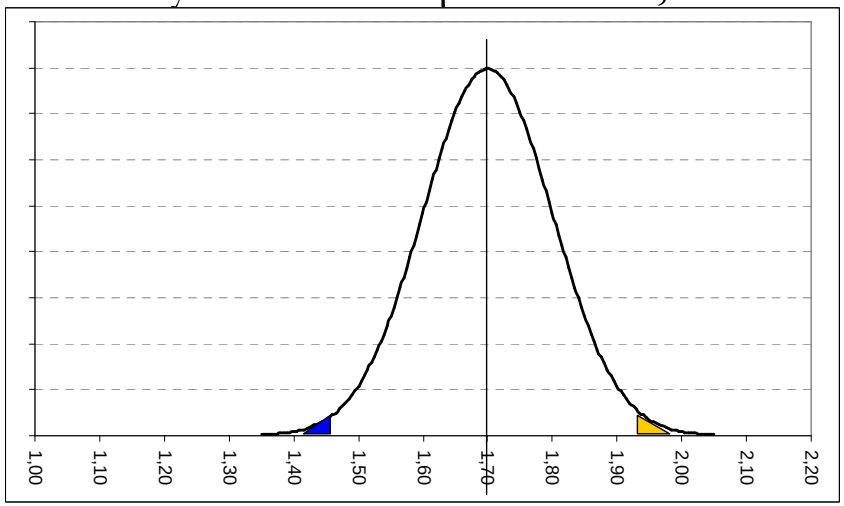




Test unilatéral – test bilatéral

- Plutôt que de regarder un seuil donné comme 5%, on peut se calculer la probabilité exacte de se tromper en rejetant H0 pour un $m_{empirique}$
- On trouve par exemple un $m_{empirique}$ =1, 94
- Dans le cadre d'un test unilatéral, on procède classiquement : on calcule la surface représentée par la partie orange qui commence au point 1,94.
- Dans le cadre du test bilatéral, on considère que la surface jaune ne représente que la moitié de la probabilité de se tromper et on « ajoute » la partie bleue qui est en miroir de l'autre côté de la courbe. En effet on calcule la probabilité que l'écart absolu à la moyenne soit supérieur à 0,24.





Test bilatéral ou test unilatéral ?

- Le test bilatéral plus exigeant pour rejeter H0.
  - En général :

    $P(H0_{bilatérale}\ vraie) = 2 * P(H0_{unilatérale}\ vraie)$

    Si on peut rejeter H0 au seuil de 5% pour un test unilatéral, on ne peut rejeter qu'au seuil de 10% dans le cadre d'un test bilatéral.
  - La pratique conduit généralement à favoriser l'hypothèse nulle d'égalité plutôt que celle d'inégalité (même si on a une inégalité en tête).
  - Raison exigence d'une part + détecter toutes sortes d'écarts au-delà de ceux préconçus.
- NB : Le test unilatéral plus exigeant pour « accepter » H0.

Le test de Student

- Test pour tester les égalités et les inégalités de moyennes
- Très utile à connaître. C'est le même test qui sert pour tester la nullité des paramètres dans une régression.

Rappels

- Moyenne : Somme des valeurs des observations divisée par l'effectif

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

- <u>Calcul sur une population de la variance et de l'écart- type :</u>
  - Variance est un indicateur de dispersion. C'est la moyenne des carrés des écarts à la moyenne.

$$\sigma^2 = \sum_{i=1}^{n} \frac{\left(X_i - \overline{X}\right)^2}{n}$$

  - L'écart-type est aussi un indicateur de dispersion plus pratique que la variance car plus en rapport avec la moyenne. C'est la racine carrée de la variance

$$\sigma = \sqrt{\sum_{i=1}^{n} \frac{\left(X_i - \overline{X}\right)^2}{n}}$$

Rappels suite

- <u>Sur un échantillon :</u> Pour des raisons de convergence, on utilise une variance et un écart-type légèrement modifiés. Au lieu de diviser par n on divise par n-1

- Variance : Moyenne des carrés des écarts à la moyenne

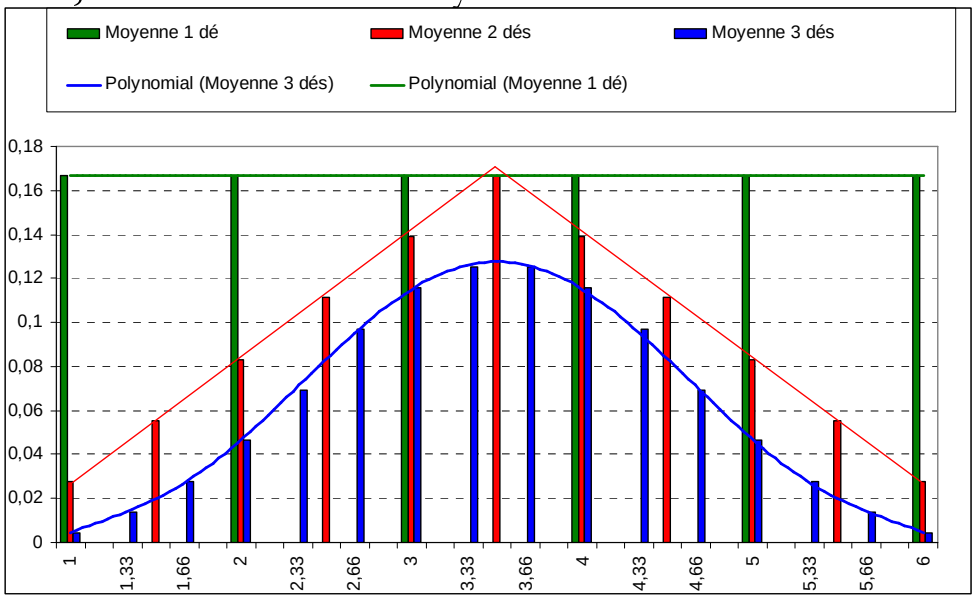
$$s^2 = \sum_{i=1}^{n} \frac{\left(X_i - \overline{X}\right)^2}{n-1}$$

- Ecart-type : Racine carrée de la variance

$$s = \sqrt{\sum_{i=1}^{n} \frac{\left(X_i - \overline{X}\right)^2}{n-1}}$$

Fondements du test de student.

- Distribution d'échantillonage = La distribution des moyennes des échantillons de taille N
- Théorème Central Limite (TCL) : La distribution d'échantillonage se rapproche de la loi normale au fur et à mesure que N augmente (et ce quelle que soit la forme de la distribution de départ).
- Exemple le dé à 6 faces, loi uniforme. Moyenne de 3 dés --> distribution en cloche...



Le test de student (suite)

- On fait des hypothèses théoriques pour H0 : sous H0, la moyenne pour la population est μ et l'écart-type σ.
- Le TCL nous dit que la distribution des moyennes d'échantillons d'une taille n se rapproche de la loi normale
- On se dote d'une loi proche de la loi normale : la loi de Student
- En tirant des échantillons de même taille que l'échantillon empirique dans une loi de student de paramètres H0, on regarde la probabilité d'engendrer des écarts à H0 au moins aussi grands que ceux observés empiriquement entre les paramètres empiriques et ceux de l'hypothèse H0.

Les éléments du test de student quand on compare un échantillon et des paramètres théoriques d'une population (μ,σ)

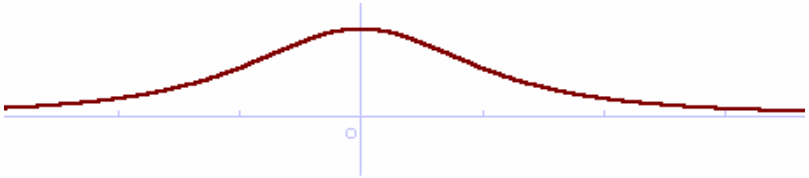
Une mesure particulière des écarts: la statistique de student *t*

$$t = \frac{\left|\overline{X} - \mu\right|}{\sigma\sqrt{\frac{1}{N}}}$$

Le degré de liberté :

DL=*N-1*

La loi de student : Graphe plus aplati que la loi normale, s'en rapprochant quand DL augmente. Formule de densité particulièrement barbare.



Densité de la loi de Student à deux degrés de liberté

$$f_T(t) = \frac{1}{\sqrt{k\pi}} \frac{\Gamma(\frac{k+1}{2})}{\Gamma(\frac{k}{2})} \frac{1}{(1+\frac{t^2}{k})^{\frac{k+1}{2}}},$$

Un calcul de probabilité

- $P_{Student\ (N-1)}(X<-t \text{ et } X>t) = 2-2*T_{N-1}(t)$

- Problème 1 : on ne connaît pas l'écart-type de la population σ. On remplace alors par l'écart type empirique de l'échantillon s.
- Problème 2 : on ne s'intéresse généralement pas à l'écart à une moyenne de la population qu'on ne connaît pas mais à la différence de moyenne de deux échantillons.

Comparaison de deux moyennes

- Hypothèses :
  - Normalité de la distribution d'échantillonage
  - Égalité des variances. (En principe il faudrait le tester...)
- Mise en oeuvre
  - Calcul de la variance commune : $s_p^2 = \frac{(N_1 - 1)s_1^2 + (N_2 - 1)s_2^2}{N_1 + N_2 - 2}$
  - Statistique t : $t = \frac{\left|\overline{X_1} - \overline{X_2}\right|}{\sqrt{s_p^2\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}}$
  - Le degré de liberté : DL=N_1+N_2-2

Exemple : le nombre de chiens possédés par les ménages hommes et les ménages femmes

CSEX	Nb	Moyenne	Écart-type
Homme	4579	0,5628	0,8846
Femme	1303	0,2678	0,6292

Différence de moyenne : 0,2949

Variance commune :

$[(4579-1)*0,8846^2+(1303-1)*0,6292^2]/[4579+1303-2] = 0,6969$

Écart-type commun : $\sqrt{Vcom}=0,8349$

Erreur-type : $0,8349* \sqrt{[(1/4597)+(1/1303)]}=0,0262$

t =DifMoy/ErreurType = 11,25

DL : 4579+1303-2=5880

Proba = $2-2*T_{5880}(11,25)=4,58*10^{-29}$

Le test du chi-deux

- Usage :
  - Test student : bon pour les moyennes.
  - On peut utiliser le test de student pour comparer deux proportions…
  - Mais il existe aussi un autre outil pour les proportions (surtout quand il y en a plus de deux) et en particulier les tableaux croisés : Le Chi-deux
- Qu'est-ce qu'on teste dans un tableau croisé ?
  - On veut montrer généralement que certaines catégories sont plus ceci que telles autres (en gros que les pratiques ne sont pas les mêmes).
  - Ho : On ne teste pas à proprement parler une égalité ou une inégalité comme dans le test de student mais le problème complexe suivant : la répartition observée des réponses a la même répartition que celle d'un tableau fictif où les deux variables croisées seraient indépendantes.

Fondement du test de chi-deux

- Loi multinomiale (loi des proportions) : tirage de n boules avec remise dans une urne avec p1 boules de couleur 1, p2 boule de couleur 2, p3 de couleur 3… pm de couleur m. Probabilité d'avoir k1 boules de couleur 1, k2 de couleur 2, … km de couleur m.
- Le carré (pondéré) de l'écart entre la proportion des boules tirées d'une couleur donnée dans le tirage et leur proportion globale dans l'urne tend vers une loi de chi-deux lorsque n tend vers l'infini
- H0 pour le chi-deux : la répartition des proportions observées suit une répartition théorique abstraite donnée.
- H0 la plus commune pour un chi-deux dans un tableau croisé : la répartition des proportions observées suit la répartition des proportions d'un tableau (abstrait) où les deux phénomènes croisés sont indépendants.
- On calcule la probabilité qu'un échantillon de taille n tiré dans une urne ait un écart au moins aussi grand à la répartition théorique que celui mesuré entre l'échantillon empirique et la distribution théorique.

Tableau empirique observé et tableau à l'indépendance

- Voici notre tableau observé.

	chien	pasdechien	Total
Immeuble	458	2157	2615
Maison	1609	1647	3256
Total	2067	3804	5871

- Construisons le tableau fictif à l'indépendance

	chien	pasdechien	Total
Immeuble	35%	65%	100%
Maison	35%	65%	100%
Total	**35%**	**65%**	**100%**

	chien	pasdechien	**Total**
Immeuble	45%	45%	**45%**
Maison	55%	55%	**55%**
Total	100%	100%	**100%**

- Taux de possession de chien et habitat sont ici indépendants si le taux de possession de chien ne dépend pas l'habitat.
- C'est le donc tableau où sur chaque ligne on retrouve le même pourcentage en ligne que le pourcentage en ligne total ou de manière équivalente, celui où on trouve sur chaque colonne le % en colonne total.

	chien	pasdechien	Total
Immeuble	=2615*35%	=45%*3804	2615
Maison	=3256*2067/5871	=3804*3256/5817	3256
Total	2067	3804	5871

- En appliquant aux effectifs marginaux (produit du % en ligne total avec l'effectif total en ligne ou de manière équivalente du % en colonne total à l'effectif total en colonne), on trouve que l'indépendance est bien le produit des marges divisé par l'effectif total.

Formule pour des tableaux croisés

- Statistique (ou distance) du chi-deux : Somme des carrés des écarts entre l'effectif observé et l'effectif théorique, pondérés par l'effectif théorique.

$$D^2 = \sum_{i=1}^{p}\sum_{i=1}^{q} \frac{\left(n_{ij} - \frac{n_{i.}n_{.j}}{n}\right)^2}{\frac{n_{i.}n_{.j}}{n}} = n\sum_{i=1}^{p}\sum_{i=1}^{q} \frac{\left(p_{ij} - p_{i.}p_{.j}\right)^2}{p_{i.}p_{.j}}$$

DL = (p-1)*(q-1) où p nombre de lignes et q nombre de colonnes (sans compter les lignes et les colonnes de total ou d'intitulés !)

- Le test du chi-deux est un test unilatéral (et non généralement bilatéral comme le test de student)
- $P_{\text{Chi-deux}[(p-1)*(q-1)]}(X>D^2)=1-\chi^2_{(p-1)(q-1)}[D]$

Mise en oeuvre simple sous Excel

- 1. Tableau des effectifs observés
- 2. Tableau des effectifs théoriques à l'indépendance : produit des marges divisé par l'effectif
- 3. Tableau des écarts entre effectifs observés et théoriques
- 4 . Tableau des écarts au carré pondérés par l'effectif théorique
- 5. Statistique du chi-deux : D^2= Somme du tableau 4
- 6. Degré de liberté: dl = (p-1)*(q-1)
- 7. On calcule grâce à la loi de probabilité du chi-deux, la probabilité d'être supérieur à la statistique D^2 avec un tel degré de liberté. Dans la cellule :

  =Loi.Khideux (D^2; dl)
- C'est la probabilité de se tromper si on rejette l'hypothèse H0 d'indépendance. Si cette probabilité est petite <10%. On peut prendre ce risque.

Exemple sous SAS

Fréquence Requis(e) Écart Cell. Khi-2 Pourcentage Pourct. en ligne Pourct. en col.	chien	pasdechien	Total
Immeuble	458	2157	2615
	920.66	1694.3	
	-462.7	462.66	
	232.5	126.34	
	7.80	36.74	44.54
	17.51	82.49	
	22.16	56.70	
Maison	1609	1647	3256
	1146.3	2109.7	
	462.66	-462.7	
	186.73	101.46	
	27.41	28.05	55.46
	49.42	50.58	
	77.84	43.30	
Total	2067	3804	5871
	35.21	64.79	100.00

Fréquence manquante = 11

Statistiques pour table de ti_b par txchien_b

Statistique	DF	Valeur	Proba.
Khi-2	1	647.0328	<.0001

Test de Kolmogorov

- Test sur deux distributions
  - Non paramétrique (pas d'hypothèse sur la distribution d'échantillonage ni sur les variances)
- On ordonne X*i* et Y*i* en fonction des valeurs X et Y. On calcule les fonctions de répartition F(Xi) et G(Yi)
- D=max|F(Xi)-G(Yi)|
- Pour n1 et n2 < 40. Cf. Tables
- Pour n1 et n2 > 40. On refuse H0 si

$$D > c.\sqrt{\frac{N_1 + N_2}{N_1.N_2}}$$

Test de Wilcoxon ou Mann Whitney

- Voici par exemple deux échantillons de taille 10.

  5,7 3,2 8,4 4,1 6,9 5,3 1,7 3,2 2,5 7,4

  8,1 5,5 3,4 7,9 4,6 1,6 8,5 7,1 8,7 5,7

- Voici les statistiques d'ordre de l'échantillon de taille 20 regroupées (les valeurs Xi du premier échantillon sont soulignées).

  1,6 <u>1,7</u> <u>2,5</u> <u>3,2</u> <u>3,2</u> 3,4 <u>4,1</u> 4,6 <u>5,3</u> 5,5

  5,7 <u>5,7</u> <u>6,9</u> 7,1 <u>7,4</u> 7,9 8,1 <u>8,4</u> 8,5 8,7

- La statistique Wx c'est la somme des rangs du plus petit échantillon

  2+3+4+5+7+9+12+13+15+18 = 88

- $U=W_x-n_x.(n_x+1)/2 = \sum_{i=1}^{n_x}\sum_{j=1}^{n_y} 1_{X_i>Y_j}$

Conclusion

- Les tests reposent fortement sur l'effectif.
  - Plus l'effectif est important, plus on a confiance dans nos données
  - Quand l'effectif est très important (millions), toute différence, même infime est significative...
  - Biais de publication
  - Biais d'estimation

  Savoir juger des tests et des critères de significativité avec doigté en tenant compte de ce fait !!!