

Test du Khi-deux

Le test d'adéquation du χ^2 est un test statistique utilisé pour vérifier si un échantillon de données suit une distribution théorique donnée (uniforme sur $\{0, 1, 2, \dots, 9\}$, Poisson de paramètre 3, loi de Poisson, loi géométrique, etc.).

1 Présentation du test

1.1 hypothèses et principe

Hypothèse nulle H_0 : « Les données observées suivent la distribution théorique spécifiée »

Hypothèse alternative H_1 : « Les données ne suivent pas la distribution théorique »

Tout le principe du test est de garantir qu'à la fin, en répondant qu'on accepte ou qu'on rejette H_0 , on aura :

$$P_{H_0}(\text{rejeter } H_0) < \alpha$$

Autrement dit, le rejet de H_0 ne se fait pas à la légère. Par exemple, on se prémunit de ce risque car le rejet de H_0 entraîne un danger pour la santé, une remise en cause d'un procédé industriel, la mise au rebut de votre générateur de nombres aléatoires, une dépense très importante, etc.

On va calculer

$$V = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i}$$

Sachant H_0 , V est proche de 0. On raisonne comme suit :

- Si V est trop grand (V dépasse un seuil qu'on présentera après), alors H_0 est très probablement fausse. On rejettera H_0 .
- Sinon, on l'accepte.

1.2 étapes du test

1. Définir les classes.

Diviser les données en k classes (intervalles, valeurs).

Exemple : Pour tester l'uniformité de nombres entre 0 et 99, on peut créer 10 classes $[0, 9]$, $[10, 19]$, ... $[90, 99]$.

2. Calculer les effectifs observés (O_i) et théoriques (E_i).

O_i : nombre d'observations dans chaque classe

E_i : nombre attendu sous H_0

3. Calculer la statistique V .

$$V = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i}$$

Nous admettons que pour n « grand », V suit la loi du χ^2 à d degrés de liberté, où $d = k - 1 - p$ où p est le nombre de paramètres estimés. Dans notre exemple d'une loi uniforme sur $[[0, 9]]$, ce sera juste $d = 9$.

4. Calculer par ailleurs avec Python le seuil associé au risque choisi.

Le risque est α et la confiance est $1 - \alpha$ (par exemple, 95%).

```
1 from scipy.stats import chi2
2 d = 9 # degrés de liberté
3 for alpha in [0.01, 0.05, 0.1]: # différents risques
4     seuil = chi2.ppf(1 - alpha, d)
5     print(seuil)
```

renvoie les seuils 21.7, 16.9, 14.7.

ppf est l'abréviation de percent point function.

risque α	0	0,01	0,05	0,1
confiance	100%	99%	95%	90%
seuil ($d = 9$)	$+\infty$	21,7	16,9	14,7

Si on n'accepte pas de se tromper, c'est qu'on veut $P_{H_0}(\text{rejeter } H_0) = 0$, donc on ne rejette jamais H_0 ...

5. Décider.

Si $V > \text{seuil}$, on rejette H_0 et on conclut que les données ne suivent pas la distribution théorique.

Sinon, on ne rejette pas H_0 .

1.3 exemple concret

Testons si un générateur de nombres aléatoires produit des entiers uniformément distribués entre 0 et 9. On génère $n = 1000$ nombres.

Classe	0	1	2	3	4	5	6	7	8	9
O_i	95	105	98	102	99	101	97	103	100	100
E_i	100	100	100	100	100	100	100	100	100	100

Calcul de la statistique :

$$V = \frac{(95 - 100)^2}{100} + \frac{(105 - 100)^2}{100} + \dots + \frac{(100 - 100)^2}{100} = 1,9$$

Degrés de liberté : $d = 10 - 1 = 9$.

Valeur seuil au risque 0.05 : 16.9.

Comme $1,9 \leq 16,9$, on ne rejette pas H_0 .

1.4 conditions d'application

- Le nombre de données doit être assez grand (1000, 10000 si possible).
- 1 les observations doivent être indépendantes.
- 1 1 il faut que $E_i \geq 5$ pour chaque classe. Si ce n'est pas le cas, il faut regrouper des classes.

Bibliographie

- Cours : Polycopié de probabilités de l'Université de Lyon (section sur les tests d'adéquation, https://math.univ-lyon1.fr/irem/IMG/pdf/PolyTunis_A_Perrut.pdf).
- Et pour les générateurs de nombres aléatoires : The Art of Computer Programming, vol. 2, Donald Knuth.

Cas particulier du test de l'uniformité pour générateurs de nombres aléatoires ; idées supplémentaires de Donald Knuth, pages 46, 47, 48)

Ce paragraphe est pertinent pour tester des générateurs de nombres aléatoires. On travaille à l'aide de la table suivante, qui peut être obtenue comme plus haut avec Python, et qui donne

$$P(V \leq x) \simeq p \text{ où } V \hookrightarrow \chi^2(\nu)$$

x étant l'entrée en ligne ν et colonne p . On rappelle que pour n grand, la statistique $\sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i}$ suit la loi $\chi^2(d)$ où $d = k - 1 - (\text{nombre de paramètres estimés})$.

Table 1
SELECTED PERCENTAGE POINTS OF THE CHI-SQUARE DISTRIBUTION

	$p = 1\%$	$p = 5\%$	$p = 25\%$	$p = 50\%$	$p = 75\%$	$p = 95\%$	$p = 99\%$
$\nu = 1$	0.00016	0.00393	0.1015	0.4549	1.323	3.841	6.635
$\nu = 2$	0.02010	0.1026	0.5754	1.386	2.773	5.991	9.210
$\nu = 3$	0.1148	0.3518	1.213	2.366	4.108	7.815	11.34
$\nu = 4$	0.2971	0.7107	1.923	3.357	5.385	9.488	13.28
$\nu = 5$	0.5543	1.1455	2.675	4.351	6.626	11.07	15.09
$\nu = 6$	0.8721	1.635	3.455	5.348	7.841	12.59	16.81
$\nu = 7$	1.239	2.167	4.255	6.346	9.037	14.07	18.48
$\nu = 8$	1.646	2.733	5.071	7.344	10.22	15.51	20.09
$\nu = 9$	2.088	3.325	5.899	8.343	11.39	16.92	21.67
$\nu = 10$	2.558	3.940	6.737	9.342	12.55	18.31	23.21
$\nu = 11$	3.053	4.575	7.584	10.34	13.70	19.68	24.72
$\nu = 12$	3.571	5.226	8.438	11.34	14.85	21.03	26.22
$\nu = 15$	5.229	7.261	11.04	14.34	18.25	25.00	30.58
$\nu = 20$	8.260	10.85	15.45	19.34	23.83	31.41	37.57
$\nu = 30$	14.95	18.49	24.48	29.34	34.80	43.77	50.89
$\nu = 50$	29.71	34.76	42.94	49.33	56.33	67.50	76.15
$\nu > 30$	$\nu + \sqrt{2\nu}x_p + \frac{2}{3}x_p^2 - \frac{2}{3} + O(1/\sqrt{\nu})$						
$x_p =$	-2.33	-1.64	-0.674	0.00	0.674	1.64	2.33

Enfin, on rappelle que n doit être suffisamment grand pour que les E_i dépassent 5. Il faudra prendre plusieurs gammes de valeurs de n ($n = 1000$, $n = 10000$, mais aussi des valeurs un peu plus petites, qui permettront peut-être de détecter des comportements non aléatoires sur de plus petites séries ; en effet, prendre n très grand a un effet de lissage des valeurs).

Jusqu'à présent, avec le test du Khi-deux classique, nous avons observé que des valeurs de V *anormalement* grandes conduisaient à rejeter H_0 . Mais examinons aussi les trop petites valeurs de V : par exemple, pour $\nu = 9$,

$$P(V \leq 2.088) \simeq 0,01$$

Observer une valeur de la statistique V égale à 0,14, par exemple, pour 9 degrés de liberté est rarissime : les valeurs O_i observées sont tellement proches des E_i qu'on ne peut plus considérer qu'elles relèvent du hasard.

Voici donc comment Donald Knuth nous suggère de procéder :

1. Faire le calcul de V sur 3 à 9 jeux de données différents :
 - Si V est inférieure à l'entrée 1% ou supérieure à l'entrée 99%, rejeter immédiatement H_0 .
 - Si V est entre les entrées 1% et 5% ou entre les entrées 95% et 99%, le générateur est *suspect*.
 - Si V est entre les entrées 5% et 10% ou entre les entrées 90% et 95%, le générateur est *presque-suspect*.
2. Décider.
 - Si au moins 2/3 des jeux de données sont "*suspect*", alors il faut considérer que le générateur n'est pas suffisamment bon.
 - Sinon, si au moins 2/3 des jeux de données sont "*suspect*" ou "*presque-suspect*", alors le générateur est limite ; il faut mener d'autres investigations.