Recherche documentaire de tests d'adéquation

TRAN Quoc Nhat Han - Amine MOUSTATIH

12 octobre 2018

Sommaire

Résumé

Ce rapport est un petit recherche documentaire concernant quelques tests d'adéquation usuels :
— Test du χ^2
— Test de Shapiro-Wilk
— Test d'Anderson Darling

1 Test χ^2

1.1 Histoire en bref

Le test du χ^2 (prononcé *khi deux* ou *khi carré*) fournit une méthode pour déterminer la nature d'une répartition, qui peut être continue ou discrète.

1.2 Processus

À la base d'un test de statistique classique, il y a la formulation d'une hypothèse appelée hypothèse nulle (ou hypothèse zéro), notée H_0. Elle suppose que les données considérées proviennent de variables aléatoires suivant une loi de probabilité donnée, et l'on souhaite tester la validité de cette hypothèse.

1. On répartit les valeurs de l'échantillon (de taille n) dans k classes distinctes et on calcule les effectifs de ces classes. Il faut vérifier que pour les i de 1 à k, les effectifs théoriques dans chacune des classes soient au moins égal à 5 (éventuellement répartir les valeurs autrement). Appelons o_i $(i = 1, ..., k)$ les effectifs observés et e_i les effectifs théoriques.
2. Calculer
$$Q = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$$
La statistique Q donne une mesure de l'écart existant entre les effectifs théoriques attendus et ceux observés dans l'échantillon. En effet, plus Q sera grand, plus le désaccord sera important. La coïncidence sera parfaite si $Q = 0$.
3. On compare ensuite cette valeur Q avec une valeur $\chi^2_{k-1,\alpha}$ issue d'un tableau (voir extrait ci-contre) à la ligne $k - 1$ et à la colonne a. $\nu = k - 1$ est le nombre de degrés de liberté et a la tolérance.
4. Si $Q > \chi^2_{k-1,\alpha}$, et si n est suffisamment grand, alors l'hypothèse d'avoir effectivement affaire à la répartition théorique voulue est à rejeter avec une probabilité d'erreur d'au plus α.

1.3 Exemple

On a lancé un dé 90 fois et on a obtenu les issues 1 à 6 ($k = 6$) avec les effectifs suivants : $12, 16, 20, 11, 13, 18$ (on a vérifié que 90 lancers sont suffisants : $n * \frac{1}{6} * \frac{5}{6} \geq 5$ implique que $n \geq 36$).

ν \ α	0,99	0,975	0,95	0,90	0,10	0,05	0,025	0,01	0,001
1	0,0002	0,001	0,004	0,016	2,71	3,84	5,02	6,63	10,83
2	0,02	0,05	0,10	0,21	4,61	5,99	7,38	9,21	13,82
3	0,11	0,22	0,35	0,58	6,25	7,81	9,35	11,34	16,27
4	0,30	0,48	0,71	1,06	7,78	9,49	11,14	13,28	18,47
5	0,55	0,83	1,15	1,61	9,24	11,07	12,83	15,09	20,51
6	0,87	1,24	1,64	2,20	10,64	12,59	14,45	16,81	22,46
7	1,24	1,69	2,17	2,83	12,02	14,07	16,01	18,48	24,32
8	1,65	2,18	2,73	3,49	13,36	15,51	17,53	20,09	26,12
9	2,09	2,70	3,33	4,17	14,68	16,92	19,02	21,67	27,88
10	2,56	3,25	3,94	4,87	15,99	18,31	20,48	23,21	29,59

FIGURE 1 – Table de la loi χ^2 pour $K \sim \chi^2(\nu)$ et α tel que $P(K \geq \chi_{\nu,\alpha})^2 = \alpha$

Si le dé n'est pas pipé (notre hypothèse), on attend comme effectifs moyens théoriques 15 pour toutes les issues.

$$Q = \frac{(12-15)^2}{15} + \frac{(16-15)^2}{15} + \frac{(20-15)^2}{15} + \\ + \frac{(11-15)^2}{15} + \frac{(13-15)^2}{15} + \frac{(18-15)^2}{15} = \frac{64}{15} = 4,266$$

Pour $\nu = k - 1 = 5$ degrés de liberté et un seuil de tolérance de 5%, la valeur $\chi_{\nu,\alpha}$ du tableau (1) est $11,07$. Cela signifie que la probabilité que Q soit supérieur à $11 \geq 1$ est de 5% (voir figure (3) ci-dessous). Comme $4,266 < 11,1$, on accepte l'hypothèse selon laquelle le dé est régulier.

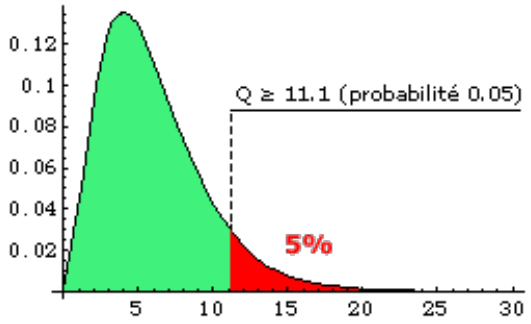

FIGURE 2 – Fonction de répartition de la loi du khi deux pour 5 degrés de libertés

1.4 Test MATLAB

1. Créez un objet de la distribution de probabilité normale standard. Générer un vecteur x de 100 valeurs selon de la distribution.

   ```
   pd = makedist('Normal');
   x = random(pd,100,1);
   ```

2. Tester l'hypothèse nulle que les données en x proviennent d'une population avec une distribution normale.

   ```
   h = chi2gof(x)
   ```

   `chi2gof(x)` retourne une décision de test de l'hypothèse nulle que les données de vecteur x provient d'une distribution normale avec une moyenne et la variance estimée à partir de x, en utilisant le test du χ^2 d'ajustement. L'autre hypothèse est que les données ne provient pas d'une telle distribution. Le résultat de h est 1 si le test rejette l'hypothèse nulle au seuil de signification de 5% et 0 sinon.

```
pd =

  NormalDistribution

  Normal distribution
       mu = 0
    sigma = 1


h1 =

     0
```

FIGURE 3 – Résultat de test χ^2 en MATLAB

2 Test de Shapiro-Wilk

2.1 Histoire en bref

Ce test non-paramétrique est publié en 1965 par *Samuel Sanford Shapiro* et *Martin Wilk* pour tester si un échantillon d'une variable **continue** (sous 2000 observations) suit **une loi normale**.

2.2 Processus

Soit une variable continue X dont n observations étaient réalisées $x_1, x_2, ..., x_n$.
Soient 2 hypothèses :
— H_0 : X suit une loi normale $N(\mu, \sigma^2)$.
— H_1 : X ne suit pas la loi normale.
Nous effectuons le test comme suivant :

1. Ordonner les réalisations dans l'ordre croissant $y_1 \leq y_2 \leq ... \leq y_n$.
2. Calculer
$$S^2 = \sum_1^n (y_i - \overline{y})^2 = \sum_1^n (x_i - \overline{x})^2$$
Donc $\overline{y}, \overline{x}$ désignent la moyenne de y, x respectivement.
3. Calculer des différences (entre le premier y_1 et le dernier y_n, le deuxième y_2 et l'avant-dernier y_{n-1}, et ainsi la suite, le médian y_{k+1} est ignoré si $n = 2k + 1$). Appliquer un coefficient de pondérer lu dans la table 4. Les additionner et élever au carré.
$$b^2 = \left(\sum_1^{\lfloor \frac{n}{2} \rfloor} a_i \left(y_{n+1-i} - y_i \right) \right)^2$$

i \ n	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	0·7071	0·7071	0·6872	0·6646	0·6431	0·6233	0·6052	0·5888	0·5739
2	—	·0000	·1677	·2413	·2806	·3031	·3164	·3244	·3291
3	—	—	—	·0000	·0875	·1401	·1743	·1976	·2141
4	—	—	—	—	—	·0000	·0561	·0947	·1224
5	—	—	—	—	—	—	—	·0000	·0399

FIGURE 4 – Coefficients a_{n+1-i} pour $n = 1..10$

4. Calculer le statistique du test

$$W = \frac{b^2}{S^2}$$

5. Rechercher la valeur W dans la table 5. Petit W signifie une distribution non normale. Choisir la valeur plus proche à W dans la

	Level								
n	**0·01**	**0·02**	**0·05**	**0·10**	**0·50**	**0·90**	**0·95**	**0·98**	**0·99**
3	0·753	0·756	0·767	0·789	0·959	0·998	0·999	1·000	1·000
4	·687	·707	·748	·792	·935	·987	·992	·996	·997
5	·686	·715	·762	·806	·927	·979	·986	·991	·993
6	0·713	0·743	0·788	0·826	0·927	0·974	0·981	0·986	0·989
7	·730	·760	·803	·838	·928	·972	·979	·985	·988
8	·749	·778	·818	·851	·932	·972	·978	·984	·987
9	·764	·791	·829	·859	·935	·972	·978	·984	·986
10	·781	·806	·842	·869	·938	·972	·978	·983	·986

FIGURE 5 – Pourcentage de W pour $n = 1..10$

ligne correspondant à n. Regarder alors le niveau (*level*) de signifiance $p-value$. Si $p-value > \alpha$ (α est le risque, souvent 1% ou 5%), l'hypothèse H_0 est accepté.

2.3 Code en R

```
shapiro.test(x)
```

x désigne un vecteur de données.

Exemples

```
> shapiro.test(rnorm(100, mean = 5, sd = 3))

Résultat de la commande:

Shapiro-Wilk normality test

data: rnorm(100, mean = 5, sd = 3)
W = 0.9895, p-value = 0.6211
```

p est significativement plus grand que $\alpha = 5\%$. L'hypothèse nulle est donc non rejetable.

```
> shapiro.test(runif(100, min = 2, max = 4))

Résultat de la commande:
Shapiro-Wilk normality test

data: runif(100, min = 2, max = 4)
W = 0.9337, p-value = 8.077e-05
```

p est trop faible. L'échantillon, par conséquence, ne suit pas la loi normale.

3 Anderson Darling

3.1 Histoire en bref

Inventé par *Theodore Wilbur Anderson* (1918–2016) et *Donald A. Darling* (1915–2014) le 1952, ce test est pour but de vérifier si un échantillon est issu d'une loi **continu** donnée. Quoi qu'il est similaire au test Kolmogorov-Smirnov, il fait beaucoup plus attention au **queue** de la courbe de la fonction de distribution en tenant compte la loi à vérifier dans le formule. C'est pourquoi les valeurs critiques dépendent de la loi à vérifier.

Ce test peut se généraliser pour tester la vraisemblance entre une famille de distribtion (la fonction de répartition n'est pas forcément spécifiée), ou estimer les paramètres.

3.2 Processus

Soit une variable continue X dont n observations étaient réalisées $x_1, x_2, ..., x_n$.
Soient 2 hypothèses :

— H_0 : X suit une loi dont $F(x)$ est la fonction de distribution cumulative. Les paramètres de F sont connus.

— H_1 : X ne suit pas la loi normale.

Nous effectuons le test comme suivant :

1. Ordonner les réalisations dans l'ordre croissant $y_1 \leq y_2 \leq ... \leq y_n$.
2. Calculer le statistique du test

$$A^2 = -n - \sum_{i=1}^{n} \frac{2i-1}{n} \left(\ln F\left(Y_i\right) + \ln\left(1 - F\left(Y_{n+1-i}\right)\right)\right)$$

3. Comparer A avec les valeurs critiques de la loi à vérifier. Si A est supérieur, l'hypothèse H_0 est rejeté. Autrement, on regarde le $p - value$.

3.3 Code en R

Il faut utiliser le package *nortest*.

```
library(nortest)
ad.test(x)
```

x est le vecteur de données.

Exemples

```
set.seed(403)
y1 = rnorm(n, mean = 0, sd = 1)
ad.test(y1)
```

```
>         Anderson-Darling normality test
>
> data:  y1
> A = 0.3093, p-value = 0.5568
```

Avec $\alpha = 0.05 < p - value$, nous acceptons H_0.

```
set.seed(403)
y2 = rexp(n,rate=1) - rexp(n,rate=1)
ad.test(y2)
```

```
>         Anderson-Darling normality test
>
> data:  y2
> A = 13.6652, p-value < 2.2e-16
```

Avec $\alpha = 0.05 > p - value$, nous rejetons H_0.

Références

[1] S. S. Shapiro et M. B. Wilk, *An analysis of variance test for normality (complete samples)*
https://github.com/haghish/ST516/blob/master/Articles/Shapiro-Wilks%20test/An%20analysis%20of%20variance%20test%20for%20normality%20(complete%20samples).pdf

[2] http://www.jybaudot.fr/Inferentielle/testsnormalite.html

[3] http://www.sthda.com/french/wiki/test-de-normalite-avec-r-test-de-shapiro-wil

[4] https://stat.ethz.ch/R-manual/R-patched/library/stats/html/shapiro.test.html

[5] https://en.wikipedia.org/wiki/Anderson%E2%80%93Darling_test

[6] https://itl.nist.gov/div898/handbook/eda/section3/eda35e.htm

[7] Christophe Chesneau, *Tables de valeurs* https://chesneau.users.lmno.cnrs.fr/tables-valeurs.pdf