Statistiques inférentielles
5 - Tests d'adéquation, d'indépendance, du khi-deux

A. BOURHATTAS

CY-Tech ING2-GSI

Année universitaire 2023-2024

Tests d'adéquation

Introduction :

Nous allons maintenant découvrir quelques test non paramétriques, en commençant par ceux permettant de répondre à la question :

- Est-ce que notre échantillon suit une loi usuelle (donnée)?
- Echantillon gaussien? Loi uniforme? Loi de Poisson?..., etc.
- Nous verrons que l'on peut utiliser des méthodes graphiques ou des tests plus classiques.

QQ-plot :

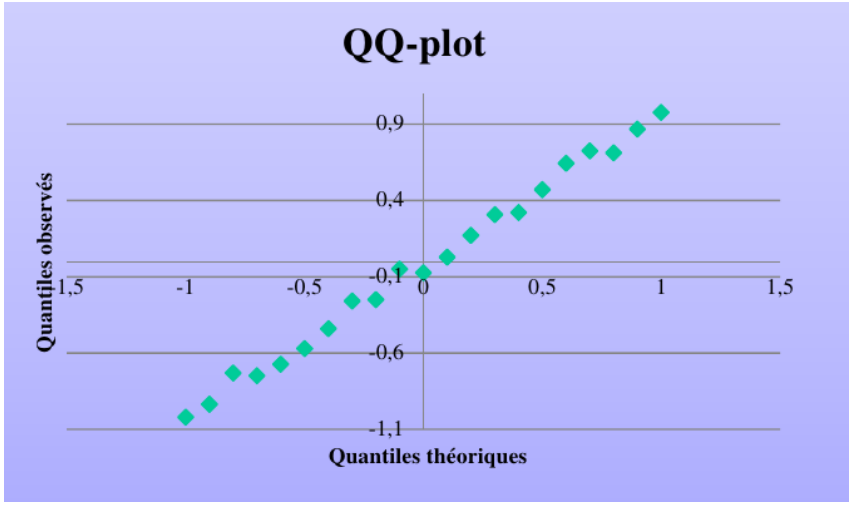
- Pour tester si un échantillon donné suit une loi usuelle donnée,
- on construit un estimateur de la fonction de répartition :

$$F_n(x) = \frac{Card\{i,\ t.q.\ x_i \leq x\}}{n},$$

- on calcule les fractiles u_i de la loi usuelle tels que : $F(u_i) = F_n(x_i)$,
- les points de coordonnées (x_i, u_i) devraient être alignés sur la 1ère bissectrice.
- le degré d'alignement de ces points indique le niveau d'adéquation de l'échantillon à la loi usuelle en question.

QQ-plot :

La plupart des langages ou logiciels statistiques proposent ce test graphique, appelé **QQ-plot** : droite quantile-quantile ou droite de Henry, selon les cas.



Mais il y a plus précis pour tester une adéquation.

Ecart entre distribution :

- Sur l'échantillon, nous calculons une distribution empirique F_n.
- Nous calculons la distribution théorique F et mesurons la distance entre les deux $d(F_n, F)$.
- Nous testerons alors : $\begin{cases} (H_0) & d(F_n, F) = 0 \\ (H_1) & d(F_n, F) \neq 0 \end{cases}$
- La région critique : $W = \{d(F, F_n) > C\}$
- si $d(F, F_n) < C$ alors l'échantillon suit la loi usuelle,
- si $d(F, F_n) > C$ alors il ne suit pas la loi usuelle,
- reste à déterminer la distance à utiliser.

Kolmogorov-Smirnov :

Il s'agit d'un test d'adéquation à une loi continue dont on connait la fonction de répartition F.

- La distance considérée est : $d(F_n, F) = Max|F_n(x_i) - F(x_i)|$.
- Le seuil, et donc la région critique sont déterminés à partir d'une table qui se présente sous la forme :

$n \backslash^{\alpha}$	0.001	0.01	0.02	0.05	0.1	0.15	0.2
1		0.99500	0.99000	0.97500	0.95000	0.92500	0.90000
2	0.97764	0.92930	0.90000	0.84189	0.77639	0.72614	0.68377
3	0.92063	0.82900	0.78456	0.70760	0.63604	0.59582	0.56481
4	0.85046	0.73421	0.68887	0.62394	0.56522	0.52476	0.49265
5	0.78137	0.66855	0.62718	0.56327	0.50945	0.47439	0.44697
6	0.72479	0.61660	0.57741	0.51926	0.46799	0.43526	0.41035
7	0.67930	0.57580	0.53844	0.48343	0.43607	0.40497	0.38145
8	0.64098	0.54180	0.50654	0.45427	0.40962	0.38062	0.35828
9	0.60846	0.51330	0.47960	0.43001	0.38746	0.36006	0.33907

Khi-deux (adéquation) 1 :

- Pour tester l'adéquation à une loi discrète,
- on part de la distribution empirique des effectifs observés $n_{i,obs}$ selon les différentes modalités,
- on calcule les effectifs théoriques $n_{i,theo}$ à partir de l'effectif total et de la loi théorique,
- dans le cas d'effectifs $n_i < 5$, on doit regrouper des modalités,
- la distance utilisée est celle du khi-deux :

$$D_n = \sum_{i=1}^{n} \frac{(n_{i,obs} - n_{i,theo})^2}{n_{i,theo}}$$

Khi-deux (adéquation) 2 :

- Distance du khi-deux :

$$D_n = \sum_{i=1}^{n} \frac{(n_{i,obs} - n_{i,theo})^2}{n_{i,theo}}.$$

- Sous l'hypothèse (H_0), D_n suit la loi du khi-deux à $(k-1)$ d.d.l. avec $k =$ nombre de modalités retenues.
- $D_n \sim \chi^2_{k-1}$
- La lecture de la table donne le seuil recherché,
- On peut alors comparer D_n au seuil ou bien lire la p-valeur pour conclure.

Khi-deux (adéquation) exemple :

Comment vérifier, à partir de 60 tirages, si un dé est équilibré ou bien truqué ?

- Voici les tirages :

Modalité	x_i	1	2	3	4	5	6
Effectifs obs	$n_{i,obs}$	11	8	9	12	7	13
Effectifs théo	$n_{i,theo}$	10	10	10	10	10	10
distance	$\frac{(n_{i,obs}-n_{i,theo})^2}{n_{i,theo}}$	0.1	0.4	0.1	0.4	0.9	0.9

- On obtient : $d_n = 2.8$.
- La table de χ^2_5 donne, pour $\alpha = 0.05$, un seuil de $C = 11.07$
- $d_n < C$, on valide (H_0).
- Ce dé est bien équilibré.

Tests d'indépendance de variables qualitatives

Khi-deux (indépendance) 1 :

- Pour tester l'indépendance de deux vars qualitatives X et Y,
- on part du tableau de contingence, c'est à dire le relevé des effectifs croisés : $n_{ij} = \text{Effectif}((X = x_i) \cap (Y = y_j))$.

$X \setminus Y$	y_1	y_2	...	y_q	Total
x_1	n_{11}	n_{12}	...	n_{1q}	$n_{1\bullet}$
x_2	n_{21}	n_{22}	...	n_{2q}	$n_{2\bullet}$
...	...	...	...	...	...
x_p	n_{p1}	n_{p2}	...	n_{pq}	$n_{p\bullet}$
Total	$n_{\bullet 1}$	$n_{\bullet 2}$	...	$n_{\bullet q}$	$n_{\bullet\bullet}$

- Les sommes par ligne $n_{i\bullet}$ et les sommes par colonne $n_{\bullet j}$ donnent la distribution marginale de Y et de X.
- $n_{\bullet\bullet}$ est l'effectif total.

Khi-deux (indépendance) 2 :

- Le but est de tester : $\begin{cases} (H_0) & X \text{ et } Y \text{ indépendants} \\ (H_1) & X \text{ et } Y \text{ liés} \end{cases}$
- la distribution théorique, en cas d'indépendance, est obtenue grâce à la formule :
  $P((X = x_i) \cap (Y = y_j)) = P((X = x_i) \times P((Y = y_j)$.
- Au niveau des effectifs, cela donne :

$$n_{ij}^{th} = \frac{n_{i\bullet} \times n_{\bullet j}}{n_{\bullet\bullet}} = \frac{\text{total ligne } i \times \text{total colonne } j}{\text{effectif total}}.$$

- La distance entre les deux distributions est alors calculée par la formule :

$$D_n = \sum_{i,j} \frac{(n_{ij}^{th} - n_{ij}^{obs})^2}{n_{ij}^{th}}$$

Khi-deux (indépendance) 3 :

- Distance du khi-deux :

$$D_n = \sum_{i,j} \frac{(n_{ij}^{th} - n_{ij}^{obs})^2}{n_{ij}^{th}}.$$

- Sous l'hypothèse (H_0), D_n suit la loi du khi-deux à $(p-1)(q-1)$ d.d.l. avec $p =$ nombre de modalités de X et $q =$ nombre de modalités de Y.
- $D_n \sim \chi^2_{(p-1)(q-1)}$
- La lecture de la table donne le seuil recherché,
- On peut alors comparer D_n au seuil ou bien lire la p-valeur pour conclure.

Remarque importante :

- L'utilisation des tests du khi-deux exige que les effectifs théoriques de chaque modalité soient supérieurs à 5.
- Dans le cas contraire, il faut fusionner des modalités jusqu'à garantir le respect de cette condition.
- Pour les test d'adéquation du khi-deux, lorsque certains paramètres de la loi théorique ne sont pas connus, mais estimés, on doit en tenir compte dans le nombre de degrés de liberté.
- Dans ce cas, on enlève un d.d.l. par paramètre estimé pour la loi du khi-deux utilisée.