13 juin 2025

1 Test du Chi2

Un essai clinique sur 200 personnes, dont 92 ont été soumises au traitement évalué, a mis en évidence que 84 d'entre elles n'ont plus de symptômes après une semaine de traitement. 90 des personnes non traitées n'ont plus de symptômes après une semaine non plus.

On cherche à déterminer si le traitement est efficace.

1. **Tables de contingence**
   (a) Établir la table de contingence observée correspondant à ces données. Quelle proportion de personnes traitées guérissent ? Quelle proportion de personnes non traitées guérissent ? Notre but sera de déterminer si cette différence est significative.
   (b) Estimer la probabilité p qu'une personne soit traitée. Estimer la probabilité q qu'une personne guérisse (indépendamment du traitement).
   (c) Supposer que le traitement n'a aucun effet. Quelle serait alors la table de contingence ?
   (d) Interpréter la distance du chi2 de la table de contingence observée (cf section 2.2.1 du poly) comme une distance entre la table de contingence observée (a) et la table de contingence théorique (c).

Soient $Y_1, Y_2, \ldots, Y_k$ k variables aléatoires réelles iid, suivant une gaussienne standard. On pose

$$Z_k = \sum_{i=1}^{k} Y_i^2.$$

On dit que Z_k suit une loi du chi2 à k degrés de liberté. On note $Z_k \sim \chi^2_k$. (Cette loi vous a déjà été présentée dans les exercices de Probabilités II.) Le tableau 1 donne la valeur de $\mathbb{P}(Z_k > z)$ pour quelques valeurs de k et de z.

On admettra[1] la proposition suivante : Soient deux variables aléatoires réelles X et Y indépendantes, ayant respectivement chacune K et L modes. Soit n la taille d'un échantillon aléatoire de (X,Y) et d_{χ^2} la distance du chi2 de la table de contingence de cet échantillon. Alors quand $n \to +\infty$,

$$d_{\chi^2} \xrightarrow{\mathcal{L}} Z_{(K-1)(L-1)}.$$

2. **Test du chi2**
   (a) Proposer un test statistique (hypothèses, statistique de test, région critique) permettant de tester l'hypothèse selon laquelle le traitement est efficace.
   (b) Que peut-on dire de notre traitement sous $\alpha = 10\%$? $\alpha = 1\%$?

1. La question 3 de ce problème permet de démontrer cette propriété dans le cas où on compare les proportions observées d'une variable à deux modes aux proportions attendues.
Pour une preuve, on pourra se reporter à l'article *Seven proofs of the Pearson Chi-squared independence test and its graphical interpretation* par É. Benhamou et V. Melot (2018), https://arxiv.org/abs/1808.09171.

(c) **Fraude scientifique.** À un niveau de signification de 5%, à combien de personnes traitées faudrait-il trouver une bonne raison pour les exclure de l'étude afin de pouvoir rejeter l'hypothèse nulle et affirmer le succès du test ?

Ce test s'appelle le test d'indépendance du chi2, et est implémenté dans `scipy.stats` :

```
import scipy.stats as st
st.chi2_contingency(np.array([[a00, a01], [a10, a11]]), correction=False)
```

3. **Loi du chi2**

   (a) Quelle sont l'espérance et la variance de Z_k ?

   (b) Soit $n \in \mathbb{N}^*, 0 < 1 < p_0$, et N_0 une variable aléatoire qui suit une loi binômiale de paramètres n et p_0 : N est la somme de n variables aléatoires réelles iid dont la loi est une loi de Bernoulli de paramètre p_0, et modélise le nombre de succès parmi n tirages d'une telle variable de Bernoulli. Posons $N_1 = n - N_0$ et $p_1 = 1 - p_0$. Montrer que quand $n \to +\infty$,

$$\frac{(N_0 - np_0)^2}{np_0} + \frac{(N_1 - np_1)^2}{np_1} \xrightarrow{\mathcal{L}} Z_1.$$

	.995	.990	.975	.950	.900	.100	.050	.025	.010	0.005	0.002	0.001
1	0.00	0.00	0.00	0.00	0.02	2.71	3.84	5.02	6.63	7.88	9.55	10.83
2	0.01	0.02	0.05	0.10	0.21	4.61	5.99	7.38	9.21	10.60	12.43	13.82
3	0.07	0.11	0.22	0.35	0.58	6.25	7.81	9.35	11.34	12.84	14.80	16.27
4	0.21	0.30	0.48	0.71	1.06	7.78	9.49	11.14	13.28	14.86	16.92	18.47
5	0.41	0.55	0.83	1.15	1.61	9.24	11.07	12.83	15.09	16.75	18.91	20.52

TABLEAU 1 – Table du χ^2 : Valeur de z telle que $\mathbb{P}(Z_k > z) = \alpha$ pour plusieurs valeurs de α et pour $Z_k \sim \chi^2_k$.

2 Test de Neyman-Pearson du rapport de vraisemblance

On considère un vecteur de $n \in \mathbb{N}^*$ variables aléatoires réelles $X_{1:n} := (X_1, \ldots, X_n)$, toutes indépendantes et de même loi qu'une variable gaussienne X d'espérance $\mu \in \mathbb{R}$ et de variance 1. Elles forment un échantillon aléatoire dont on observe une réalisation $x_{1:n} := (x_1, \ldots, x_n)$. On souhaite réaliser le test simple de l'hypothèse $H_0 : \mu = 0$ contre $H_1 : \mu = m$ pour un certain $m < 0$ fixé.

1. **Statistique de test**

   (a) Rappeler l'expression de la vraisemblance $L : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}$ de l'échantillon $x_{1:n}$.

   (b) Calculer le rapport de vraisemblance

$$\Lambda(x_{1:n}) = \frac{L(x_{1:n}, m)}{L(x_{1:n}, 0)}.$$

   En donner une interprétation.

   (c) Expliciter la loi de la variable aléatoire $\lambda(X_{1:n}) = \ln \Lambda(X_{1:n})$. En déduire celle de $\Lambda(X_{1:n})$. Quelles sont-elles sous les hypothèses nulle puis alternative ?

2. **Région de rejet**
   (a) En utilisant $\Lambda(X_{1:n})$ comme statistique de test, quelle est la forme de la région de rejet $\mathcal{I}_\alpha$ pour un niveau $\alpha \in]0,1[$ quelconque ?
   (b) On note Φ la fonction de répartition de la loi normale centrée-réduite, et q_α son quantile de niveau α. Exprimer $\mathcal{I}_\alpha$ en fonction de q_α.

3. **Puissance du test**
   (a) Pour $m < 0$ fixé, quelle est la puissance $\pi_n(m)$ de ce test ?
   (b) Pour n fixé, étudier les variations de $m \in \mathbb{R}_-^* \mapsto \pi_n(m)$, la fonction puissance du test. Interpréter.