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, . . . ,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.

	.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 $Z_k \sim \chi^2_k$.

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}

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.

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 $\alpha = 10\%$? $\alpha = 1\%$?
   (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 $n \in \mathbb{N}^*$, $0 < p_0 < 1$, et N_0 une variable aléatoire qui suit une loi binômiale de paramètres n et p_0 : N_0 est la somme de n variables aléatoires réelles iid dont la loi est une loi de Bernouilli de paramètre p_0, et modélise le nombre de succès parmi n tirages d'une telle variable de Bernouilli. Posons $N_1 = n - N_0$ et $p_1 = 1 - p_0$. Montrer que quand $n \to +\infty$,

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

Solution

Question 1.a

	Pas de guérison	Guérison	Total
Pas de traitement	$A_{00} = 18$	$A_{01} = 90$	$N_{0.} = 108$
Traitement	$A_{10} = 8$	$A_{11} = 84$	$N_{1.} = 92$
	$N_{.0} = 26$	$N_{.1} = 174$	$n = 200$

La proportion de personnes traitées qui guérissent est $\frac{A_{11}}{N_{1.}} = 91\%$. La proportion de personnes non-traitées qui guérissent est $\frac{A_{01}}{N_{0.}} = 83\%$.

Cette différence semble élevée. Mais l'est-elle vraiment ?

Question 1.b On peut modéliser par des Bernoulli et estimer par maximum de vraisemblance :

$$p = \frac{N_{1.}}{n} = 0.46 \text{ et } q = \frac{N_{.1}}{n} = 0.87.$$

Question 1.c Si le traitement n'a aucun effet, alors $\mathbb{P}(\text{guérir \& traitement}) = \mathbb{P}(\text{guérir})\mathbb{P}(\text{traitement})$ et on s'attend à la table de contingence suivante, pour 100 personnes :

	Pas de guérison	Guérison	Total
Pas de traitement	$B_{00} = n(1-p)(1-q) = 14$	$B_{01} = n(1-p)q = 94$	$N_{0.} = 108$
Traitement	$B_{10} = np(1-q) = 12$	$B_{11} = npq = 80$	$N_{1.} = 92$
	$N_{.0} = 26$	$N_{.1} = 174$	$n = 200$

Question 1.d La distance du chi2 est donnée par

$$\begin{aligned}
d_{\chi^2} &= \sum_{i=0}^{1}\sum_{j=0}^{1} \frac{\left(A_{ij} - \frac{N_{i.}N_{.j}}{n}\right)^2}{\frac{N_{i.}N_{.j}}{n}} \\
&= \frac{\left(A_{00} - \frac{N_{0.}N_{.0}}{n}\right)^2}{\frac{N_{0.}N_{.0}}{n}} + \frac{\left(A_{01} - \frac{N_{0.}N_{.1}}{n}\right)^2}{\frac{N_{0.}N_{.1}}{n}} + \frac{\left(A_{10} - \frac{N_{1.}N_{.0}}{n}\right)^2}{\frac{N_{1.}N_{.0}}{n}} + \frac{\left(A_{11} - \frac{N_{1.}N_{.1}}{n}\right)^2}{\frac{N_{1.}N_{.1}}{n}} \\
&= \frac{(A_{00} - n(1-p)(1-q))^2}{n(1-p)(1-q)} + \frac{(A_{01} - n(1-p)q)^2}{n(1-p)q} + \frac{(A_{10} - np(1-q))^2}{np(1-q)} + \frac{(A_{11} - npq)^2}{npq} \\
&= \frac{(A_{00} - B_{00})^2}{B_{00}} + \frac{(A_{01} - B_{01})^2}{B_{01}} + \frac{(A_{10} - B_{10})^2}{B_{10}} + \frac{(A_{11} - B_{11})^2}{B_{11}} \\
&= \sum_{i=0}^{1}\sum_{j=0}^{1} \frac{(A_{ij} - B_{ij})^2}{B_{ij}},
\end{aligned}$$

où A_{ij} est la valeur *observée* dans la case i,j tandis que B_{ij} est la valeur *attendue* dans la case i,j.

Ainsi d_{χ^2} mesure à quel point les cases de la table de contingence observée divergent de la table que l'on observerait si les variables étaient indépendantes.

Question 2.a On propose alors le test suivant, pour n grand :

- $\mathcal{H}_0$: le traitement n'a aucun effet.
- $\mathcal{H}_1$: le traitement a un effet.
- Statistique de test : d_{χ^2}.
- Distribution de la statistique de test sous $\mathcal{H}_0$: à peu près (n grand) une chi2 à 1 degré de liberté.

Question 2.b Dans nos données,

$$d_{\chi^2} = \frac{(18-14)^2}{14} + \frac{(90-94)^2}{94} + \frac{(8-12)^2}{12} + \frac{(84-80)^2}{80} = 2.85.$$

D'après le tableau 1, la valeur critique pour $\alpha = 0.1$ est $z_{0.10} = 2.71 : \mathbb{P}(Z_1 > 2.71) = 0.1$. Nous pouvons rejeter $\mathcal{H}_0$.

Par contre, pour $\alpha = 0.01$, la valeur critique est $z_{0.01} = 6.63$. Nous ne pouvons pas rejeter $\mathcal{H}_1$ avec un niveau de signification de 1%.

Question 2.c Gardons A_{00}, A_{01} et A_{11} fixés. Comment la statistique de test évolue-t-elle quand on change A_{10} ? Numériquement (voir tableau 2), on obtient une statistique de test supérieure à $z_{0.05} = 3.84$ pour $A_{10} = 6$. Il suffit de trouver une justification à l'élimination de deux patients de l'étude pour que ses résultats semblent en devenir significatifs ($p < 0.05$).

Question 3.a L'espérance de Z_k vaut

$$\mathbb{E}(Z_k) = \sum_{i=1}^{k} \mathbb{E}(Y_i^2) \text{ par indépendance des} Y_i$$

et $\mathbb{E}(Y_i^2) = \mathbb{V}(Y_i) + \mathbb{E}(Y_i)^2$ par définition de la variance. Comme $\mathbb{E}(Y_i) = 0$ et $\mathbb{V}(Y_i) = 1$ on obtient

$$\mathbb{E}(Z_k) = k.$$

La variance de Z_k est donnée par $\mathbb{V}(Z_k) = \mathbb{E}(Z_k^2) - \mathbb{E}(Z_k)^2$. On a $\mathbb{E}(Z_k)^2 = k^2$ et

$$\begin{aligned}
\mathbb{E}(Z_k^2) &= \mathbb{E}\left(\sum_{i=1}^{k} Y_i^2 \sum_{j=1}^{k} Y_j^2\right) \\
&= \sum_{i=1}^{k}\sum_{j\neq i} \mathbb{E}(Y_i^2)\mathbb{E}(Y_j^2) + \sum_{i=1}^{k} \mathbb{E}(Y_i^4) \text{ par linéarité de l'espérance + indépendance des } Y_i \\
&= k(k-1) + 3k \text{ (cf. formule pour les moments d'une loi normale.)}
\end{aligned}$$

Ainsi $\mathbb{V}(Z_k) = 2k$.

Question 3.b $N_0 \sim \mathcal{B}(n,p_0)$ est une somme de n variables de Bernouilli d'espérance p_0 et de variance $p_0(1-p_0)$.

Par le théorème central limite,

$$\frac{N_0 - np_0}{\sqrt{np_0(1-p_0)}} \xrightarrow{\mathcal{L}} Y, \text{ où } Y \sim \mathcal{N}(0,1).$$

Donc

$$\frac{(N_0 - np_0)^2}{np_0(1-p_0)} \xrightarrow{\mathcal{L}} Z_1.$$

Enfin,

$$\begin{aligned}
\frac{(N_0 - np_0)^2}{np_0(1-p_0)} &= \frac{(N_0 - np_0)^2}{np_0(1-p_0)}(1 - p_0 + p_0) \\
&= \frac{(N_0 - np_0)^2}{np_0} + \frac{(N_0 - np_0)^2}{n(1-p_0)} \\
&= \frac{(N_0 - np_0)^2}{np_0} + \frac{(N_0 - np_0 + n - n)^2}{n(1-p_0)} \\
&= \frac{(N_0 - np_0)^2}{np_0} + \frac{(-N_1 + np_1)^2}{np_1} \qquad \square
\end{aligned}$$

```python
# Données fixées
a00 = 18
a01 = 90
a11 = 84

# Calcul de la statistique de test en fonction de a10
def compute_chi2(a10):
    n = a00 + a01 + a10 + a11
    p = float(a11 + a10)/n
    q = float(a01 + a11)/n
    b00 = (n * (1-p) * (1-q)) # int(n * (1-p) * (1-q))
    b01 = (n * (1-p) * q) # int(n * (1-p) * q)
    b10 = (n * p * (1-q)) # int(n * p * (1-q))
    b11 = (n * p * q) # int(n * p * q)
    chi2 = float((a00 - b00)**2)/b00 + float((a01 - b01)**2)/b01 + \
      float((a10 - b10)**2)/b10 + float((a11 - b11)**2)/b11
    return chi2

# Calcul de la valeur de la statistique jusqu'à dépasser le seuil voulu
for a10 in np.arange(9, 0, -1):
    chi2 = compute_chi2(a10)
    if chi2 > 3.84:
        print("a10 = %d, Chi2 = %.3f" % (a10, chi2))
        break for a10 in np.arange(8):
    print("a10 = \%d, Chi2 = \%.3f" \% (a10, compute_chi2(a10)))
```

TABLEAU 2 – Code Python pour évaluer la statistique de test du chi2 en fonction de A_{10} et déterminer la valeur maximale de A_{10} pour laquelle la statistique de test est supérieure au seuil à 0.05%.

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.

Solution

On note $\mathbf{1}_n$ le vecteur colonne de taille n dont toutes les composantes valent 1, et I_n la matrice identité de taille $n \times n$.

Question 1.a Comme vu dans l'exercice 2 de la PC 1, le vecteur aléatoire $X_{1:n}$ est ici supposé suivre une loi Normale multivariée d'espérance $\mu\mathbf{1}_n$ et de variance I_n, donc la vraisemblance de l'échantillon $x_{1:n}$ sous le modèle considéré et pour un paramètre d'espérance $\theta \in \mathbb{R}$ vaut

$$L(x_{1:n},\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{(x_i-\theta)^2}{2}\right\} = (2\pi)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\theta)^2\right\}.$$

Question 1.b On déduit de la question précédente le rapport de vraisemblance :

$$\begin{aligned}\Lambda(x_{1:n}) = \frac{L(x_{1:n},m)}{L(x_{1:n},0)} &= \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\left((x_i-m)^2 - x_i^2\right)\right\} \\ &= \exp\left\{m\sum_{i=1}^{n} x_i - \frac{n}{2}m^2\right\}\end{aligned}$$

Ce rapport compare les vraisemblances de l'échantillon sous l'hypothèse nulle (au numérateur) et l'hypothèse alternative (au dénominateur). On s'attend à ce qu'il soit d'autant plus petit que H_0 est vraisemblable par rapport à H_1, et d'autant plus grand que H_1 est vraisemblable par rapport à H_0. Il devrait avoisiner 1 si les deux hypothèses sont difficiles à distinguer, i.e. sont tout aussi vraisemblables l'une que l'autre.

Question 1.c La variable aléatoire $\lambda(X_{1:n})$ s'écrit

$$\lambda(X_{1:n}) = \ln \Lambda(X_{1:n}) = m\sum_{i=1}^{n} X_i - \frac{n}{2}m^2.$$

Sous le modèle posé, il s'agit d'une combinaison linéaire de variables aléatoires gaussiennes indépendantes, elle suit donc elle-même une loi Normale, d'espérance

$$\mu_\lambda = \mathbb{E}\left(\lambda(X_{1:n})\right) = m\sum_{i=1}^{n}\mathbb{E}(X_i) - \frac{n}{2}m^2 = n\mu m - \frac{n}{2}m^2$$

et de variance

$$\sigma_\lambda^2 = \mathbb{V}\left(\lambda(X_{1:n})\right) = m^2 \sum_{i=1}^{n} \mathbb{V}(X_i) = nm^2.$$

La variable aléatoire $\Lambda(X_{1:n}) = e^{\lambda(X_{1:n})}$ suit donc une loi log-normale de paramètres μ_λ et σ_λ^2.

Seul μ_λ change selon l'hypothèse adoptée. Sous l'hypothèse nulle il vaut

$$\mu_\lambda = -\frac{n}{2}m^2 = -\frac{1}{2}\sigma_\lambda^2,$$

tandis que sous l'hypothèse alternative il devient

$$\mu_\lambda = nm^2 - \frac{n}{2}m^2 = \frac{n}{2}m^2 = \frac{1}{2}\sigma_\lambda^2.$$

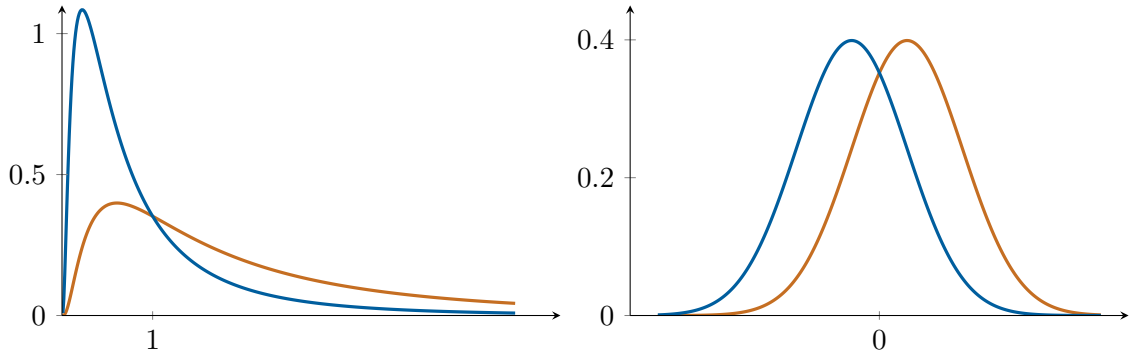

FIGURE 1 – Densités de $\Lambda(X_{1:n})$ (à gauche) et $\lambda(X_{1:n})$ (à droite) sous $H_0 : \mu = 0$ (bleu) et sous $H_1 : \mu = 1$ (orange) pour un échantillon de taille $n = 1$

Question 2.a Étant donnée l'interprétation du rapport de vraisemblance $\Lambda(x_{1:n})$ formulée à la question 1.b, la région de rejet au niveau α devrait prendre la forme $\mathcal{I}_\alpha =]\ell_\alpha, +\infty[$ où $\ell_\alpha \in \mathbb{R}_+$ est tel que $\mathbb{P}\left(\Lambda(X_{1:n}) > \ell_\alpha \mid H_0\right) = \alpha$.

Question 2.b Soit $Z \sim \mathcal{N}(0,1)$. Rappelons que

$$\begin{aligned}\ln \Lambda(X_{1:n}) &= m \sum_{i=1}^{n} X_i - \frac{n}{2}m^2 \\ &= \sqrt{n}m \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i \right) - \frac{n}{2}m^2.\end{aligned}$$

Ainsi, sous l'hypothèse nulle, $\ln \Lambda(X_{1:n})$ a la même loi que $\sqrt{n}mZ - \frac{n}{2}m^2$ et comme $m < 0$, on a :

$$\begin{aligned}\mathbb{P}\left(\Lambda(X_{1:n}) > \ell_\alpha \mid H_0\right) &= \mathbb{P}\left(\ln \Lambda(X_{1:n}) > \ln(\ell_\alpha) \mid H_0\right) \\ &= \mathbb{P}\left(\sqrt{n}mZ - \frac{n}{2}m^2 > \ln(\ell_\alpha)\right) \\ &= \Phi\left(\frac{\ln(\ell_\alpha)}{\sqrt{n}m} + \frac{\sqrt{n}}{2}m\right).\end{aligned}$$

Il vient :

$$\begin{aligned} & \Phi\left(\frac{\ln(\ell_\alpha)}{\sqrt{n}m} + \frac{\sqrt{n}}{2}m\right) = \alpha \\ \Leftrightarrow \quad & \frac{\ln(\ell_\alpha)}{\sqrt{n}m} + \frac{\sqrt{n}}{2}m = \Phi^{-1}(\alpha) = q_\alpha \\ \Leftrightarrow \quad & \ell_\alpha = \exp\left\{\sqrt{n}mq_\alpha - \frac{n}{2}m^2\right\}. \end{aligned}$$

Question 3.a La puissance du test est, par définition, la probabilité de rejeter l'hypothèse nulle à raison, soit :

$$\pi_n(m) = \mathbb{P}\left(\Lambda(X_{1:n}) \in \mathcal{I}_\alpha \mid H_1\right).$$

Reprenons :

$$\begin{aligned} \ln \Lambda(X_{1:n}) &= m\sum_{i=1}^{n} X_i - \frac{n}{2}m^2 \\ &= \sqrt{n}m\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(X_i - m)\right) + \frac{n}{2}m^2. \end{aligned}$$

Ainsi, sous l'hypothèse alternative, $\ln \Lambda(X_{1:n})$ a la même loi que $\sqrt{n}mZ + \frac{n}{2}m^2$ et comme $m < 0$, on a :

$$\begin{aligned} \mathbb{P}\left(\Lambda(X_{1:n}) > \ell_\alpha \mid H_1\right) &= \mathbb{P}\left(\ln \Lambda(X_{1:n}) > \ln(\ell_\alpha) \mid H_1\right) \\ &= \mathbb{P}\left(\sqrt{n}mZ + \frac{n}{2}m^2 > \ln(\ell_\alpha)\right) \\ &= \Phi\left(\frac{\ln(\ell_\alpha)}{\sqrt{n}m} - \frac{\sqrt{n}}{2}m\right) \\ &= \Phi\left(q_\alpha - \sqrt{n}m\right) \\ &= \Phi\left(-q_{1-\alpha} - \sqrt{n}m\right) = 1 - \Phi\left(q_{1-\alpha} + \sqrt{n}m\right) \end{aligned}$$

par symétrie de la loi normale centrée-réduite.

Question 3.b La fonction puissance $\pi_n : m \in \mathbb{R}_-^* \mapsto 1 - \Phi\left(q_{1-\alpha} + \sqrt{n}m\right)$ est strictement décroissante, avec $\lim_{m\to-\infty} \pi_n(m) = 1$ et $\lim_{m\to 0} \pi_n(m) = \alpha$. En d'autres termes, plus on considère une hypothèse alternative $\mu = m$ éloignée de l'hypothèse nulle $\mu = 0$, plus il est aisé de les distinguer et plus le test est puissant. En revanche, lorsque les deux hypothèses se rapprochent, i.e. $m \to 0$, plus il devient difficile de distinguer les lois sous chacune d'entre elles, et la puissance décroît jusqu'à atteindre le niveau du test.