---

1 Estimateur du nombre de boules dans une urne

On considère une urne contenant n boules, numérotées de 1 à n. On ne connaît pas n et on souhaite l'estimer. Pour cela, on procède à m tirages avec remise.

1. Donner un estimateur simple de n.
2. Calculer l'estimateur par maximum de vraisemblance de n.
3. Montrer que le biais de cet estimateur est

$$-\sum_{k=1}^{n-1}\left(\frac{k}{n}\right)^m.$$

Que vaut-il quand $m = 1$? Quand $m \to +\infty$?

Solution

Modélisation : Échantillon aléatoire $(X_1, X_2, \ldots, X_m)$ où X_i est le numéro de la boule tirée au i-ème tirage.

La loi de X est donnée par :

$$\mathbb{P}_X(X = k|n) = \begin{cases} \frac{1}{n} & \text{si } k \in [1, \ldots, n] \\ 0 & \text{sinon} \end{cases} = \frac{1}{n}\, 1_{[1,\ldots,n]}(k).$$

Question 1. Un estimateur naturel de n est

$$N_m = \max_{i=1,\ldots,m}(X_i).$$

Une autre possiblité est d'observer que X étant uniforme sur $[1, n]$, la moyenne empirique des X_i s'approche de $\frac{n+1}{2}$, et de suggérer

$$N_m = 2\frac{1}{m}\sum_{i=1}^{m} X_i - 1.$$

Question 2. La vraisemblance d'un échantillon $(x_1, x_2, \ldots, x_m)$ pour l'estimation η de n vaut

$$L(x_1, x_2, \ldots, x_m|\eta) = \prod_{i=1}^{m} \frac{1}{\eta}\, 1_{[1,\ldots,\eta]}(x_i).$$

L'estimation par maximum de vraisemblance de n est donc

$$\widehat{n}_{\text{MLE}} = \arg\max_{\eta\in\mathbb{N}^*} \prod_{i=1}^{m} \frac{1}{\eta}\, 1_{[1,\ldots,\eta]}(x_i).$$

La quantité $\prod_{i=1}^{m} \frac{1}{\eta}\, 1_{[1,\dots,\eta]}(x_i)$ est positive, et non nulle dès que toutes les indicatrices valent 1, c'est-à-dire que $x_i \leq \eta$ pour tout $i = 1, \dots, m$. Enfin, quand elle est non-nulle, elle est d'autant plus petite que η est grand. Ainsi $\widehat{n}_{\text{MLE}}$ doit être aussi petit que possible tout en majorant tous les x_i. On a donc

$$\widehat{n}_{\text{MLE}} = \max_{i=1,\dots,m}(x_i),$$

et l'estimateur par maximum de vraisemblance de n est la variable aléatoire réelle

$$\widehat{N}_{\text{MLE}} = \max_{i=1,\dots,m}(X_i).$$

Le premier des estimateurs que nous avons proposé à la question précédente est en fait l'estimateur par maximum de vraisemblance de N.

Question 3. Le biais de l'estimateur que nous avons proposé est, par définiition :

$$\text{B}(N_m) = \mathbb{E}(N_m) - n.$$

Calculons $\mathbb{E}(N_m)$:

$$\begin{aligned}
\mathbb{E}(N_m) &= \sum_{k=1}^{n} k\mathbb{P}(N_m = k) \text{ par définition} \\
&= \sum_{k=1}^{n} \mathbb{P}(N_m \geq k) \\
&= \sum_{k=1}^{n} 1 - \mathbb{P}(N_m \leq (k-1)) \\
&= n - \sum_{k=1}^{n} \left(\frac{(k-1)}{n}\right)^m \\
&= n - \sum_{k=1}^{n-1} \left(\frac{k}{n}\right)^m.
\end{aligned}$$

La deuxième ligne s'obtient en observant que

$$\begin{aligned}
\sum_{k=1}^{n} \mathbb{P}(N_m \geq k) &= \mathbb{P}(N_m \geq n) + \mathbb{P}(N_m \geq n-1) + \cdots + \mathbb{P}(N_m \geq 1) \\
&= \mathbb{P}(N_m = n) + (\mathbb{P}(N_m = n) + \mathbb{P}(N_m = n-1)) + \cdots + \\
&\quad (\mathbb{P}(N_m = n) + \mathbb{P}(N_m = n-1) + \cdots + \mathbb{P}(N_m = 1)),
\end{aligned}$$

puis en regroupant ensemble les n termes $\mathbb{P}(N_m = n)$, les $(n-1)$ termes $\mathbb{P}(N_m = n-1)$, etc.

La quatrième ligne s'obtient en observant que l'événement "$N_m \leq k-1$" est équivalent à l'événement "$X_i \leq k-1$ pour $i = 1, \dots, m$", que $\mathbb{P}(X_i \leq k-1) = \frac{k-1}{n}$, et que les X_i sont indépendants.

Ainsi notre estimateur est biaisé, est son biais vaut

$$\text{B}(N_m) = -\sum_{k=1}^{n-1} \left(\frac{k}{n}\right)^m.$$

Ce biais est négatif : la valeur estimée est en moyenne plus faible que le nombre de boules. Il est peu probable de tirer la boule numéro n, sauf à faire un très grand nombre de tirages.

Tirer une seule boule ($m = 1$) est équivalent à tirer une valeur uniformément entre 1 et n, on s'attend donc à être plus proche de $\frac{n+1}{2}$ (espérance d'une variable aléatoire réelle uniformément distribuée sur $[1, n]$) que de n. On a bien

$$\mathrm{B}(N_1) = -\sum_{k=1}^{n-1} \left(\frac{k}{n}\right) = \frac{1}{n}\frac{n(n-1)}{2} = \frac{n-1}{2} = n - \frac{n+1}{2}.$$

Par ailleurs, le biais tend vers 0 quand $m \to +\infty$. N_m est asymptotiquement non biaisé.

2 Estimation de densité

Considérons une variable aléatoire X suivant une loi normale de paramètres μ et σ^2.

1. Étant donné un échantillon de $n \in \mathbb{N}^*$ observations de X, calculer l'estimateur par maximum de vraisemblance de μ et σ.
2. Supposons maintenant que μ est la réalisation d'un variable aléatoire réelle M qui suit une loi normale de moyenne m et de variance τ^2. Calculer l'estimateur de Bayes de μ.
3. Décomposer l'estimateur de Bayes de μ en la somme d'un terme fonction de la moyenne empirique de l'échantillon et un terme dépendant de la moyenne a priori. Que se passe-t-il quand n augmente ?

Solution

Question 1 Appelons $(x_1, x_2, \ldots, x_n)$ notre échantillon. Il s'agit d'une réalisation de l'échantillon aléatoire $(X_1, X_2, \ldots, X_n)$ composé de variables i.i.d. de même loi que X. La vraisemblance de cet échantillon est

$$L(x_1, x_2, \ldots, x_n; \mu, \sigma) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right),$$

et sa log-vraisemblance est donc

$$\ell(x_1, x_2, \ldots, x_n; \mu, \sigma) = \sum_{i=1}^{n} -\ln(\sigma\sqrt{2\pi}) - \frac{(x_i-\mu)^2}{2\sigma^2}.$$

Nous cherchons

$$\widehat{\mu}_{\mathrm{MLE}},\ \widehat{\sigma}_{\mathrm{MLE}} \in \operatorname*{arg\,max}_{\widehat{\mu},\widehat{\sigma}\in\mathbb{R}^2} \ell(x_1, x_2, \ldots, x_n; \widehat{\mu}, \widehat{\sigma}).$$

La fonction $(\mu, \sigma) \mapsto \sum_{i=1}^{n} -\ln(\sigma\sqrt{2\pi}) - \frac{(x_i-\mu)^2}{2\sigma^2}$ est une fonction concave dont le gradient en μ vaut $\frac{\partial \ell}{\partial \mu} = \frac{1}{\sigma^2}\left(n\mu - \sum_{i=1}^{n} x_i\right)$. En annulant ce gradient, on obtient $\widehat{\mu}_{\mathrm{MLE}} = \frac{1}{n}\sum_{i=1}^{n} x_i$. L'estimateur par maximum de vraisemblance de l'espérance de X est sa moyenne empirique.

Le gradient en σ de la log-vraisemblance vaut $\frac{\partial \ell}{\partial \sigma} = -\frac{n}{\sigma} + \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{\sigma^3}$. En remplaçant μ par $\widehat{\mu}_{\mathrm{MLE}}$ et en annulant ce gradient, on obtient $\widehat{\sigma}^2_{\mathrm{MLE}} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \widehat{\mu}_{\mathrm{MLE}})^2$ et $\widehat{\sigma}_{\mathrm{MLE}}$ est donc l'écart-type empirique.

Question 2 Nous supposons maintenant que μ est la réalisation d'une variable aléatoire réelle $M \sim \mathcal{N}(m, \tau^2)$. Son estimateur de Bayes est $\widehat{\mu}_{\text{Bayes}} = \mathbb{E}(M|X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$.

Pour déterminer cette espérance, calculons la densité correspondante :

$$\begin{aligned}
\mathbb{P}(M = \mu|X_1, X_2, \ldots, X_n) &= \frac{\mathbb{P}(X_1, X_2, \ldots, X_n|M = \mu)\mathbb{P}(M = \mu)}{\mathbb{P}(X_1, X_2, \ldots, X_n)} \text{ (Bayes)} \\
&= \frac{1}{\mathbb{P}(X_1, X_2, \ldots, X_n)} \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left(-\sum_{i=1}^n \frac{(X_i - \mu)^2}{2\sigma^2}\right) \frac{1}{\tau\sqrt{2\pi}} \exp\left(-\frac{(\mu - m)^2}{2\tau^2}\right) \\
&= \mathcal{K}_1 \exp\left(-\sum_{i=1}^n \frac{(X_i - \mu)^2}{2\sigma^2} - \frac{(\mu - m)^2}{2\tau^2}\right) \text{ où } \mathcal{K}_1 \text{ ne dépend pas de } \mu.
\end{aligned}$$

Ainsi,

$$\begin{aligned}
\mathbb{P}(M = \mu|X_1, X_2, \ldots, X_n) &= \mathcal{K}_1 \exp\left(-\frac{1}{2}\left[\frac{1}{\sigma^2}\sum_{i=1}^n (X_i^2 + \mu^2 - 2X_i\mu) + \frac{1}{\tau^2}(\mu^2 + m^2 - 2\mu m)\right]\right) \\
&= \mathcal{K}_1 \exp\left(-\frac{1}{2}\left[\left(\frac{n}{\sigma^2} + \frac{1}{\tau^2}\right)\mu^2 - 2\left(\frac{\sum_{i=1}^n X_i}{\sigma^2} + \frac{m}{\tau^2}\right)\mu + \frac{\sum_{i=1}^n X_i^2}{\sigma^2} + \frac{m^2}{\tau^2}\right]\right) \\
&= \mathcal{K}_1 \exp\left(-\frac{1}{2}\frac{(\mu - a)^2}{b^2}\right) \exp\left(-\frac{1}{2}\mathcal{K}_2\right),
\end{aligned}$$

où $\mathcal{K}_1$ et $\mathcal{K}_2$ ne dépendent pas de μ,

$$a = \frac{\sum_{i=1}^n X_i\tau^2 + m\sigma^2}{n\tau^2 + \sigma^2} \text{ et } b = \sqrt{\frac{\tau^2\sigma^2}{n\tau^2 + \sigma^2}}.$$

L'intégrale d'une densité de probabilité vaut 1. Cela vaut pour la loi normale centrée en a et d'écart-type b, comme pour $\mathbb{P}(M|X_1, X_2, \ldots, X_n)$, et donc $\mathcal{K}_1 \exp\left(-\frac{1}{2}\mathcal{K}_2\right) = \frac{1}{b\sqrt{2\pi}}$.

Ainsi, $M|X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$ suit une loi normale centrée en a. Son espérance vaut donc a, et ainsi

$$\widehat{\mu}_{\text{Bayes}} = \frac{\sum_{i=1}^n x_i\tau^2 + m\sigma^2}{n\tau^2 + \sigma^2}.$$

Question 3

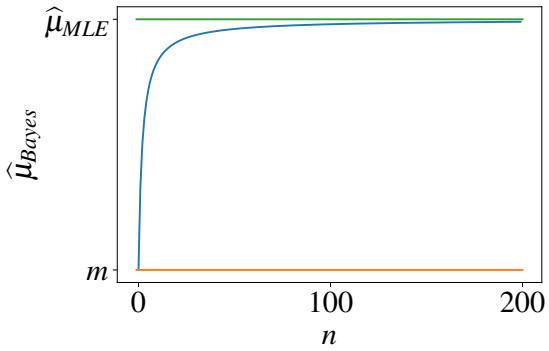
$$\widehat{\mu}_{\text{Bayes}} = \frac{n/\sigma^2}{n/\sigma^2 + 1/\tau^2}\widehat{\mu}_{\text{MLE}} + \frac{1/\tau^2}{n/\sigma^2 + 1/\tau^2}m.$$

Plus il y a de données et plus l'estimateur de Bayes est proche de la moyenne empirique. Quand il y a peu de données, l'estimateur de Bayes est plus proche de la valeur a priori du paramètre. Ce phénomène est illustré sur la figure 1.

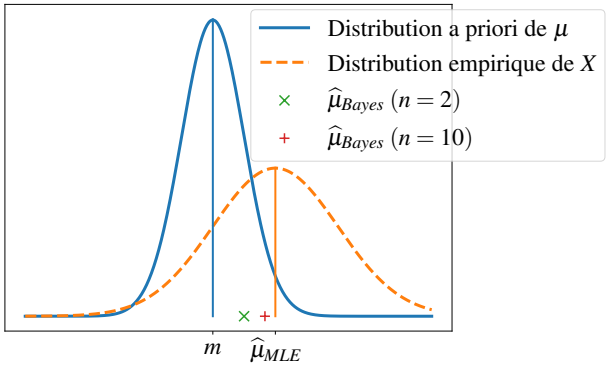
3 Test du Chi2

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.



(A) Estimation de Bayes en fonction de la taille de l'échantillon.



(B) Distribution empirique de X et distribution a priori de μ.

FIGURE 1 – Estimation de Bayes de la moyenne d'une gaussienne. Ici $\widehat{\mu}^2_{\text{MLE}} = 2$ et $\widehat{\sigma}^2_{\text{MLE}} = 2$, tandis que $m = 0$ et $\tau^2 = 1$. Plus n est grand, plus on s'éloigne de la valeur a priori de μ pour se rapprocher de son estimation empirique.

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.

	.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} 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 $n \in \mathbb{N}^*$, $0 < 1 < p_0$, 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 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 ?

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.

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 $\mathbb{P}(\text{guérir}|\text{traitement}) = \mathbb{P}(\text{guérir})\mathbb{P}(\text{traitement})$ 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

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

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 table 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,

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

```
# 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%.