Variables aléatoires - Lois de probabilité

Exercice 1 (Modèle Statistique)

Nous nous trouvons face à une urne noire contenant N boules numérotées de 1 à N. Le nombre N est inconnu, mais nous pouvons tiré autant de boules que nous le souhaitons, à condition de replacer chaque boule dans l'urne avant d'en tirer une nouvelle.

1) Écrire le modèle statistique

2) Comment peut-on estimer la valeur de N à partir des nombres $x_1, \dots, x_n$ observés après n tirages ?

3) Quelle est la distribution du maximum $\hat{N} = \max(X_1, \dots X_n)$?

4) Combien de boules souhaitez-vous tirer ? Indice : calculer la probabilité que $\hat{N} = N$.

Correction

1) Nous tirons une boule au hasard, et appelons X_1 la variable aléatoire correspondant au numéro indiqué sur la boule. Nous répétons cette expérience n fois, de sorte que nous obtenons n variables aléatoires $X_1, \dots, X_n$ où, pour $1 \leq i \leq n$, X_i correspond au nombre indiqué sur la i-ème boule tirée. Les variables $X_1, \dots, X_n$ sont indépendantes et identiquement distribuées. De plus, pour tout $1 \leq i \leq n$, et pour tout $1 \leq k \leq N$, la boule numérotée k a une probabilité $1/N$ d'être tirée en i-ème position. Ainsi, pour $1 \leq k \leq N$: $\mathbb{P}(X_i = k) = 1/N$. Ceci correspond à la loi uniforme sur $\{1, \dots, N\}$ avec N inconnu. Le modèle statistique associé à X_1 est :

$$\{\mathcal{U}(\{1, \dots, N\}); N \in \mathbb{N}\},$$

où $\mathcal{U}(\{1, \dots, N\})$ indique la distribution uniforme sur $\{1, \dots, N\}$. Comme $X_1, \dots, X_n$ sont i.i.d., le modèle statistique est le même pour chaque v.a., et le modèle pour $(X_1, \dots, X_n)$ s'écrit :

$$\{\mathcal{U}(\{1, \dots, N\})^{\otimes n}; N \in \mathbb{N}\}.$$

2) On cherche à trouver le plus grand des numéros inscrits sur les boules. Ainsi, on peut par exemple estimer N à l'aide du plus grand numéro tiré :

$$\hat{N} = \max_{1 \leq i \leq n} X_i.$$

3) Pour caractériser la distribution du maximum $\hat{N} = \max_{1 \leq i \leq n} X_i$, nous utilisons la fonction de répartition $F(x) = \mathbb{P}(x \leq \hat{N})$, $x \in \mathbb{R}$:

$$\begin{aligned}\mathbb{P}(\hat{N} \leq t) &= \mathbb{P}(X_i \leq x,\ \forall\ i,\ 1 \leq i \leq n,) \\ &= \big(\mathbb{P}(X_1 \leq x)\big)^n = \begin{cases} \left(\frac{\lfloor x \rfloor}{N}\right)^n & \text{if } 0 \leq x < N \\ 0 & \text{if } x < 0 \\ 1 & \text{otherwise.} \end{cases}\end{aligned}$$

4) Pour approcher N au mieux, on veut que la probabilité que $\hat{N} = N$, $\mathbb{P}(\hat{N} = N)$, soit la plus proche de 1 possible. Or :

$$\mathbb{P}(\hat{N} = N) = \mathbb{P}(\hat{N} \leq N) - \mathbb{P}(\hat{N} \leq N-1) = 1 - \left(\frac{N-1}{N}\right)^n \xrightarrow{n \to \infty} 1.$$

On veut donc que n soit le plus grand possible, ce qui revient à tirer le plus de boules possible.

Exercice 2 (Régression linéaire)

1) L'entreprise a un échantillon de taille $n = 25$. Il s'agit d'un échantillon d'une population hypothétique de toutes les bielles possibles. Pour chaque $i \leqslant n$, ils ont mesuré x_i le poids brut de coulée et y_i le poids final de la bielle. On suppose que $(x_1, y_1), \ldots, (x_n, y_n)$ sont des réalisations de variables aléatoires i.i.d. bivariées (X_i, Y_i) for $i \leqslant n$. En observant le nuage de points de l'échantillon, on constate que les points (x_i, y_i) ont l'air d'être alignés. On peut penser que le poids final est une fonction linéaire du poids de la coulée brute. Cela donne : $Y_i = aX_i + b$; Ce modèle conduirait à des points parfaitement alignés, ce qui n'est pas le cas en raison du bruit. Nous ajoutons donc un peu de bruit à ce modèle, et nous modélisons la relation entre les deux poids de la manière suivante : $Y_i = aX_i + b + \epsilon_i$ avec ϵ_i une variable aléatoire distribuée à partir d'une certaine distribution centrée $P\theta$, par exemple, la distribution normale $\mathcal{N}(0, \sigma^2)$. X_i est distribuée selon une distribution Q inconnue, et conditionnellement à X_i, Y_i est distribué selon $\mathcal{N}(aX_i + b, \sigma^2)$. Les paramètres décrivant le modèle sont alors $\theta = (a, b, \sigma^2) \in \mathbb{R}^2 \times \mathbb{R}_+$ et Q.
2) L'entreprise veut "quantifier la relation entre" les deux poids, c'est-à-dire qu'elle veut connaître les paramètres décrivant le modèle, c'est-à-dire estimer a, b et σ.

Exercice 3 (Tests en régression linéaire)

1) L'estimateur des moindres carrés est défini par :

$$(\hat{\beta}_0, \hat{\beta}_1) \in \operatorname{argmin} \sum_{i=1}^{1} 0(Y_i - \beta_0 - \beta_1 X_i)^2.$$

En écrivant les conditions d'optimalités du premier ordre (annulation du gradient), on obtient les valeurs

$$\hat{\beta}_0 = \bar{Y} - \frac{s_{XY}}{s_X^2}\bar{X}, \quad \beta_1 = \frac{s_{XY}}{s_X^2},$$

avec

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i,$$

$$s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2, \qquad s_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2,$$

$$s_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}), \qquad r = \frac{s_{XY}}{s_X s_Y},$$

$$s = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

2) D'après le cours, les statistiques suivantes suivent respectivement des lois du χ^2 et de Student :

$$\frac{(\hat{\beta}_0 - \beta_0)}{s\sqrt{\frac{1}{n} + \frac{\bar{X}^2}{(n-1)s_X^2}}} \sim \text{Student}(n-2), \qquad \frac{(\hat{\beta}_1 - \beta_1)}{s\sqrt{\frac{1}{(n-1)s_X^2}}} \sim \text{Student}(n-2).$$

Ainsi, en notant $t_{0.005}^{n-2}$ le quantile à 0.5% de la loi de Student à $n-2$ degrés de liberté, et

$$I_0 = [\hat{\beta}_0 - t_{0.005}^{n-2} s\sqrt{\frac{1}{n} + \frac{\bar{X}^2}{(n-1)s_X^2}}, \hat{\beta}_0 + t_{0.005}^{n-2} s\sqrt{\frac{1}{n} + \frac{\bar{X}^2}{(n-1)s_X^2}}],$$

$$I_1 = [\hat{\beta}_1 - t_{0.005}^{n-2} s\sqrt{\frac{1}{(n-1)s_X^2}}, \hat{\beta}_1 + t_{0.005}^{n-2} s\sqrt{\frac{1}{(n-1)s_X^2}}],$$

on a

$$\mathbb{P}(\beta_0 \in I_0) \geq 0.99,$$

$$\mathbb{P}(\beta_1 \in I_1) \geq 0.99.$$