Université d'Evry val d'Essonne
Master MIAGE

TD 4 - Statistiques - corrections

Exercice 1 - Elections

(a) Définissons X_i une variable aléatoire suivant une loi de Bernoulli, valant 1 si la personne i vote Erika, 0 sinon. X_i ~ Bern(0.5) ; E(X_i) = p = 0,5 ; V(X_i) = p (1 - p) = 0,25
Ensuite, $N_{Erika} = X_1 + \ldots + X_{400}$, donc E(N_{Erika}) = 400 * E(X_i) = 200 et V(N_{Erika}) = 400 * V(X_i) = 100

Le théorème central limite :

- **s'applique pour une somme de n variables indépendantes et de même loi X_i (c'est le cas)**
- **alors cette somme suit une loi normale N(n*E(X_i), n*V(X_i))**

Donc N_{Erika} ~ N(200, 100). On centre et on réduit pour se ramener à une loi Z ~ N(0,1):
p(N_{Erika} > 210) = p ((N_{Erika} - 200)/10 > (210-200)/10) = p(Z > 1) ≈ 0,16

(b) Définissons Y_i une variable aléatoire suivant une loi de Bernoulli, valant 1 si la personne i vote pour Pierre, Jeanne, ou Jerry, et 0 sinon.
Y_i ~ Bern(0.3) ; E(Y_i) = p = 0,3 ; V(Y_i) = p (1 - p) = 0,3*0,7 = 0,21
Les Y_i sont indépendants et de même loi, donc par le théorème central limite, la somme S = $Y_1 + \ldots + Y_{400}$ suit une loi normale N(400*0.3 , 400*0.21) = N(120, 84).
Donc p(S < 100) = p ($\frac{S-120}{\sqrt{84}} < \frac{100-120}{\sqrt{84}}$) = p(Z < $\frac{-20}{\sqrt{21\times 4}}$) = p(Z < $\frac{-10\times 2}{\sqrt{21}\times 2}$) = p(Z < $\frac{-10}{\sqrt{21}}$) ≈ 0,0145

Exercice 2 - Ajustement de courbe de données

(a) Vraissemblance = p(y_i | a,b,x_i,σ), et $y_i - ax_i - b \sim N(0,\sigma^2)$ donc $y_i \sim N(ax_i + b, \sigma^2)$, on écrit sa densité de probabilité:

$$p(y_i \vee a,b,x_i,\sigma) = f(y_i) = \frac{1}{\sigma\sqrt{2\pi}} \times e^{-(y_i - a x_i + b)^2/(2\sigma^2)}$$

(b) On a les données $x_i = (1,3,5)$ et $y_i = (8,2,1)$. Les données étant indépendantes,

$$p(y_1, y_2, y_3 \vee a, b, x_1, x_2, x_3, \sigma) = \prod_{i=1}^{3} p(y_i \vee a, b, x_i, \sigma)$$

En développant, on obtient $\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^3 \times e^{-[(8-a-b)^2+(2-3a-b)^2+(1-5a-b)^2]/(2\sigma^2)}$.

On a donc $\ln(p(8,2,1 \vee a, b, 1,3,5, \sigma))$

$$-3\ln(\sigma) - \frac{3}{2}\ln(2\pi) - \frac{[(8-a-b)^2+(2-3a-b)^2+(1-5a-b)^2]}{2\sigma^2}$$

$$-3\ln(\sigma)-\frac{3}{2}\ln(2\pi)-\frac{[35a^2+3b^2-38a-22b+18ab+69]}{2\sigma^2}$$

Dans le cas général, on écrirait:

$$p(y_1,\dots,y_n \vee a,b,x_1,\dots,x_n,\sigma)=\prod_{i=1}^{n} p(y_i \vee a,b,x_i,\sigma)$$

$$\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \times e^{-\sum_{i=1}^{n}(y_i-ax_i-b)^2/(2\sigma^2)} \quad -n\ln(\sigma)-\frac{n}{2}\ln(2\pi)-\frac{\sum_{i=1}^{n}(y_i-ax_i-b)^2}{2\sigma^2}$$

(c) Chercher les dérivées partielles de la vraissemblance par rapport à a et b. Chercher ensuite quand elles s'annulent, et en déduire la meilleure estimation de a et b.

$$\frac{\partial\ln(p(8,2,1 \vee a,b,1,3,5,\sigma))}{\partial a}=\frac{70a+18b-38}{-2\sigma^2}=0 \leftrightarrow 70a+18b=38$$

$$\frac{\partial\ln(p(8,2,1 \vee a,b,1,3,5,\sigma))}{\partial b}=\frac{6b+18a-22}{-2\sigma^2}=0 \leftrightarrow 18a+6b=22$$

$$\begin{cases}70a+18b=38\\ 18a+6b=22\end{cases} \Leftrightarrow \begin{cases}70a+18b=38\\ 54a+18b=66\end{cases} ¿\Leftrightarrow \begin{array}{l}16a=-28\\ a=-7/4\end{array}$$

$$18\times\frac{-7}{4}+6b=22 \Leftrightarrow b=\frac{1}{6}\times\left(\frac{88+7\times18}{4}\right)=\frac{88+70+56}{24}=\frac{214}{24}=\frac{107}{12}$$

Ainsi, la droite qui passe le mieux par les données a pour équation $y=\frac{-7}{4}x+\frac{107}{12}$.

Exercice 3 - Maximum de vraissemblance

(a) La densité de probabilité d'une loi uniforme sur [a,b] est

$$f(x_i \vee a,b)=\frac{1}{b-a}\, si\, x\in[a,b], 0\, sinon.$$

Les données sont indépendantes, donc la vraissemblance L vaut

$$L=\prod_{i=1}^{5} p(x_i \vee a,b)=\frac{1}{(b-a)^5}\, si(\forall i\, alors\, x_i\in[a,b]), sinon\, 0$$

On veut trouver a et b tels que L soit maximum. Donc on veut déjà que toutes les données soient dans [a,b] : on a une première condition a ≤ $\min(x_i)$ ≤ $\max(x_i)$ ≤ b.
Ensuite, il faut que (b-a) soit le plus petit possible.
On prendra donc a = $\min(x_i)$ = 1.2 et b = $\max(x_i)$ = 10.5

Ainsi, la distribution uniforme qui maximise la vraissemblance de ces données est la distribution uniforme sur l'intervalle [1.2, 10.5].
Note: il n'y a aucune raison que ce soit bien celle de laquelle on a tiré ces données. C'est juste celle pour laquelle notre observation de données est la plus probable.

(b) Dans le cas général, on prendra toujours a = min(observations) et b = max(observations).