Université d'Evry val d'Essonne
Master MIAGE

TD 4 - Statistiques

Exercice 1 - Elections

Pour nommer le prochain responsable du département de statistiques de l'UEVE, supposons que 50% de la population supporte Erika, 20% supporte Ruthi, et le reste est partagé entre Pierre, Jeanne, et Jerry. Un sondage interroge 400 personnes sur leur intention de vote.

(a) Utilisez le théorème central limite pour estimer la probabilité qu'au moins 52.5% des gens interrogés lors de ce sondage préfère Erika.

(b) Utilisez le théorème central limite pour estimer la probabilité que moins de 25% des gens interrogés préfère Pierre, Jeanne ou Jerry.

Données : pour Z ~ N(0,1) on a p(Z>1) = 0.16 $p(Z < \frac{-10}{\sqrt{21}}) \approx 0{,}0145$

p(Z > 1.96) = 0,025 p(Z > 2.58) = 0,005 p(Z>3.29) = 0,0005

Exercice 2 - Ajustement de courbe de données

Supposons que nous ayons des données à deux variables (dites "bivariées") sous la forme d'un ensemble de points $(x_1, y_1), \dots, (x_n, y_n)$.

On peut décider d'utiliser un modèle linéaire, c'est à dire supposer qu'il existe une relation linaire entre y et x, et donc que les données devraient reposer sur une droite. On appelle cette droite la droite de régression.

Cependant, puisque les données contiennent un peu de "bruit" aléatoire, et que notre modèle linéaire n'est probablement pas exact, ce ne sera pas le cas. Ce que l'on peut faire, c'est essayer de trouver la droite qui passera le mieux par les les points de données. Nous allons faire une "régression linéaire". Pour des données bivariées, le modèle de régression simple suppose que les x_i ne sont pas aléatoires mais que les y_i le sont, les y_i étant des observations d'une variable aléatoire $\mathbf{Y_i \sim ax_i + b + \varepsilon_i}$. Les ε_i représentent les erreurs de mesure, et sont définis comme des variables aléatoires normales d'espérance 0 et de variance σ^2 qu'on suppose connue. Nous les supposons indépendants les uns des autres.

Note : les x_i et y_i ne sont pas des variables, ce sont des données. Y_i et ε_i sont des variables aléatoires.

(a) La distribution de Y_i dépend de a, b, σ et x_i. De ces valeurs, seules a et b sont inconnues. Donnez la formule de la vraissemblance de y_i sachant les paramètres a, b, x_i et σ, notée $p(y_i \mid a,b,x_i,\sigma)$.

Astuce : on sait que $\mathbf{y_i - ax_i - b \sim N(0,\sigma^2)}$.

Note : une "vraissemblance", c'est toujours la probabilité d'observer une donnée sachant un modèle particulier.

(b) Supposons que nous ayons les données (1,8), (3,2) et (5,1). En vous basant sur le même modèle, écrivez la vraissemblance et la log-vraissemblance comme une fonction de a, b et σ.

Donnez aussi l'expression générale pour toutes données $(x_1, y_1), \dots, (x_n, y_n)$.

(c) Avec les valeurs de données de (b), trouver l'estimation de a et b qui maximise la vraissemblance, qu'on appelle *estimateur du maximum de vraissemblance*.

Exercice 3 - Maximum de vraissemblance

Supposons que nous ayons des données indépendantes { 1.2, 2.1, 1.3, 10.5, 5 }, que nous avons tiré d'une distribution uniforme sur l'intervalle [a,b].

(a) Donnez l'estimateur du maximum de vraissemblance de a et b. *Astuce: n'essayez pas de dériver la vraissemblance.*

(b) Donnez l'estimateur pour des données générales x_1, x_2, ... , x_n.