Statistiques appliquées (L3 d'économie) - Cours de Patrick Sevestre - TD 4 - Corrigé

Marc Sangnier - marc.sangnier@ens-cachan.fr

2 décembre 2007

Exercice 1

Question 1 - Méthode des moments

Dans le TD précédent, nous avons utilisé la moyenne empirique pour estimer la moyenne des loyers. En fait il s'agit d'une utilisation de la méthode des moments. En effet, la moyenne empirique est le moment d'ordre 1.

$m_1 = \frac{1}{n}\sum_{i=1}^{n} y_i$ est le moment d'ordre 1
$m_2 = \frac{1}{n}\sum_{i=1}^{n} y_i^2$ est le moment d'ordre 2
$m_k = \frac{1}{n}\sum_{i=1}^{n} y_i^k$ est le moment d'ordre k
$\mu_2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - m_1)^2$est le moment centré d'ordre 2
$\mu_k = \frac{1}{n}\sum_{i=1}^{n}(y_i - m_1)^k$est le moment centré d'ordre k

Question 2 - Méthodes du maximum de vraisemblance et des moindes carrés

Méthode du maximum de vraisemblance

On observe n réalisations d'une variable aléatoire qui obéit à une loi de paramètre θ. La probabilité d'oberver $(x_1; x_2; ...; x_n)$ dépend donc de la valeur du paramètre θ inconnu. L'idée de la méthode du maximum de vraisemblance est rechercher une valeur $\hat{\theta}$de θ qui maximise la probabilité d'avoir observé cet ensemble de réalisations de la varaible aléatoire.

Si on suppose que la variable aléatoire suit une loi continue de paramètre θ, sa densité peut s'écrire $f(x; \theta)$.

La probabilité d'observer la réalisation x_i est donc $f(x_i; \theta)$.

Si l'on fait l'hypothèse d'indépendance des observations, la probabilité d'observer l'échantillon $(x_1; x_2; ...; x_n)$, appelée vraisemblance, est égale au produit des densité de chaque obsrevation. Appelons $L(.)$ la fonction de vraisemblance :

$$L(\theta; x_1; x_2; ...; x_n) = f(x_1; \theta) * f(x_2; \theta) * ... * f(x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta)$$

L'idée sous-jacente est que le paramètre θ est tel que l'échantillon $(x_1; x_2; ...; x_n)$ soit celui des valeurs les plus probables de la variable aléatoire. Pour déterminer $\hat{\theta}$ on va

maximiser $L(.)$ par rapport à θ en résolvant l'équation suivante :

$$\frac{\partial L(\theta; x_1; x_2; ...; x_n)}{\partial \theta} = 0$$

Méthode des moindres carrés

La méthode des moindres carrés consite à estimer les paramètres d'un modèle de façon à minimiser l'écart entre les valeurs prédites par ce modèle pour une variable et les observations de cette variable.

Supposons que l'on dispose d'un échantillon regroupant n observation d'un vecteur $(x; y)$. Pour chaque observation i, on a donc un couple $(x_i; y_i)$.

Supposons que notre modèle soit décrit par une fonction $g(.)$ dépendant d'un paramètre θ inconnu. Pour une valeur donnée de $x = x_i$, le modèle prédit une valeur $\hat{y}_i = g(x_i; \theta)$ de y.

L'écart entre la valeur prédite est la valeur réalisée est donc $\varepsilon_i = y_i - \hat{y}_i$

On peut donc réécire le modèle sous la forme $y_i = g(x_i; \theta) + \varepsilon_i$.

ε est un variable aléatoire supposée d'espérance nulle, de variance σ^2 et indépendant de la variable x.

Pour déterminer le paramètre $\hat{\theta}$ qui minimise les écarts, on va définir une fonction de perte quadratique :

$$S = \sum_{i=1}^{n} [y_i - g(x_i; \theta)]^2$$

On va ensuite chercher à minimiser cette fonction par les méthodes habituelles.

Question 3 - Estimateurs du loyer moyen

Estimateur du loyer moyen par la méthode des moindres carrés

Soit y_i la valeur prise pour le loyer d'un studio. On suppose que tous les loyers ont la même espérance m et la même variance.

La modèle le plus simple pour décrire le loyer d'un studio est le suivant :

$$y_i = m + \varepsilon_i$$

Cette écriture revient à dire que chaque loyer y_i diffère de la moyenne de l'ensemble des loyers d'un terme ε_i qui lui est propre. Par construction $E(\varepsilon_i) = 0$ et la valeur prédite par le modèle est $\hat{y}_i = m$

Le paramètre que nous cherchons à estimer est m.

La fonction de perte s'écrit :

$$S = \sum_{i=1}^{n} (y_i - m)^2$$

Minimisons S en annulant sa dérivée par rapport à m. La condition du second ordre étant vérifiée.

$$\frac{\partial S}{\partial m} = -2 \sum_{i=1}^{n} (y_i - m) = 0$$

$$\iff \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} m = 0 \iff \sum_{i=1}^{n} y_i = n * m$$

$$\iff \hat{m} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

Estimateur du loyer moyen par la méthode du maximum de vraisemblance

On fait l'hypothèse que la distribution des loyer peut être représentée par une loi Normale d'espérance m est de variance σ^2. Par ailleurs, on suppose les observations des loyers de l'échantillon indépendantes. La densité d'une observation y_i s'écrit donc :

$$f(y_i; \sigma^2; m) = \frac{1}{\sqrt{2\Pi}\sigma} exp[-\frac{1}{2\sigma^2}(y_i - m)^2]$$

La fonction de vraisemenblance de l'échantillon est donc :

$$L(m; \sigma^2; y_1; y_2; ...; y_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\Pi}\sigma} exp[-\frac{1}{2\sigma^2}(y_i - m)^2]$$

Pour des raisons pratiques, on va utiliser la log-vraisemblance :

$$ln[L(.)] = \sum_{i=1}^{n} ln \left\{ \frac{1}{\sqrt{2\Pi}\sigma} exp[-\frac{1}{2\sigma^2}(y_i - m)^2] \right\}$$

$$\iff ln[L(.)] = \sum_{i=1}^{n} ln[(2\Pi\sigma^2)^{-\frac{1}{2}}] + \sum_{i=1}^{n} -\frac{1}{2\sigma^2}(y_i - m)^2$$

$$\iff ln[L(.)] = -\frac{n}{2} ln(2\Pi) - \frac{n}{2} ln(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - m)^2$$

La condition du second ordre étant vérifiée, on va annuler la dérivée de $ln[L(.)]$ par rapport à m :

$$\frac{\partial ln[L(.)]}{\partial m} = 0 \iff \frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - m) = 0$$

$$\iff \hat{m} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

Observations et remarques

On constate que les trois estimateurs de la moyenne issus des méthodes présentées ici (moments, moindres carrés et maximum de vraisemblance) sont égaux à la moyenne empirique. C'est estimateur ont donc les propriétés suivantes :

- Absence de biais (si toutes les observations ont la même espérance).
- Variance égale à $\frac{\sigma^2}{n}$ (si toutes les observations sont indépendantes et ont la même variance σ^2).
- Convergence presque sûre, en moyenne quadratique et en probabilité (si les observations sont indépendantes et identiquement distribuées).
- Efficacité (sous l'hypothèse de normalité).
- Distribution asymptotiquement normale (si les observation sont iid).
- Efficacité asymptotique (sous l'hypothèse de normalité).

Question 4 - Combinaison d'estimateurs

Soit n_1 la taille du premier échantillon. Soit n_2 la taille du second échantillon.
$n_1 + n_2 = n$

Méthode des moments

$$\hat{m} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n_1+n_2}\left\{\sum_{i=1}^{n_1} y_i + \sum_{j=1}^{n_2} y_j\right\}$$

$$\hat{m} = \frac{n_1}{n_1+n_2}\left\{\frac{1}{n_1}\sum_{i=1}^{n_1} y_i\right\} + \frac{n_2}{n_1+n_2}\left\{\frac{1}{n_2}\sum_{j=1}^{n_2} y_j\right\}$$

$$\hat{m} = \frac{n_1}{n_1+n_2}\hat{m_1} + \frac{n_2}{n_1+n_2}\hat{m_2}$$

Avec $\hat{m_1}$ et $\hat{m_2}$ les estimateurs de la moyenne calculés à partir des deux sous-échantillons par la méthode des moments.

Méthode des moindres carrés

$$\hat{m} = Min_m \sum_{i=1}^{n}(y_i - m)^2 = Min_m \sum_{i=1}^{n_1}(y_i - m)^2 + \sum_{j=1}^{n_2}(y_j - m)^2$$

La condition du premier ordre est :

$$-2\sum_{i=1}^{n_1}(y_i - m) - 2\sum_{j=1}^{n_2}(y_j - m) = 0$$

$$\sum_{i=1}^{n_1} y_i - n_1 m + \sum_{j=1}^{n_2} y_j - n_2 m = 0$$

$$n_1\left(\frac{1}{n_1}\sum_{i=1}^{n_1} y_i\right) - n_1 m + n_2\left(\frac{1}{n_2}\sum_{j=1}^{n_j} y_j\right) - n_2 m = 0$$

$$m(n_1+n_2) = n_1\hat{m_1} + n_2\hat{m_2} \Longleftrightarrow \hat{m} = \frac{n_1}{n_1+n_2}\hat{m_1} + \frac{n_2}{n_1+n_2}\hat{m_2}$$

Avec $\hat{m_1}$ et $\hat{m_2}$ les estimateurs de la moyenne calculés à partir des deux sous-échantillons par la méthode des moindres carrés.

Méthode du maximum de vraisemblance

En décomposant comme précédemment, on obtient la fonction de log-vraisemblance :

$$ln[L(.)] = -\frac{n_1}{2}ln(2\Pi) - \frac{n_1}{2}ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n_1}(y_i - m)^2 - \frac{n_2}{2}ln(2\Pi) - \frac{n_2}{2}ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{j=1}^{n_2}(y_j - m)^2$$

La condition du premier ordre s'écrit donc :

$$\sum_{i=1}^{n_1}(y_i - m) + \sum_{j=1}^{n_2}(y_j - m) = 0$$

On reconnaît ici l'une des étapes de la méthode des moindres carrés, ma solution sera donc la même.

Conditions de regroupement des estimateurs

On peut combiner des estimateurs comme nous l'avons fait ici et dans le TD précédent si les deux échantillons ont la même espérance, la même variance et sont indépendants.

Question 5 - Estimateurs de la variance

Méthode des moments

L'estimateur de la variance par la méthode des moments est la variance empirique, c'est à dire le moment centré d'ordre 2.

$$\hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

Avec $\bar{y}$ la moyenne empirique, c'est à dire le moment d'ordre 1 de la distribution.
Rappelez-vous que cet estimateur est biaisé.

Méthode du maximum de vraisemblance

Reprenons l'expression précédente de la log-vraisemblance :

$$ln[L(.)] = -\frac{n}{2}ln(2\Pi) - \frac{n}{2}ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - m)^2$$

$$\frac{\partial ln[L(.)]}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(y_i - m)^2 = 0$$

$$\iff \frac{1}{\sigma^2}\left([\frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - m)^2] - n\right) = 0$$

$$\iff \hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

Même remarque que pour l'estimateur construit avec la méthode des moments.

Méthode des moindres carrés

Usuellement, on définit l'estimateur de la variance par la méthode des moindres carrés par l'estimateur sans biais $\hat{\sigma^2} = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$

Question 6 - Intervalle de confiance

Rappel : Principe de construction d'un intervalle de confiance

Soit $\hat{\theta}$ un estimateur d'un paramètre θ. On désire construire un intrevalle qui contienne θ avec un probabilité donnée $1 - \alpha$.

On sait que la valeur de $\hat{\theta}$ dépend de θ. Il existe donc un intervalle contenant $\hat{\theta}$ avec un probabilité $1 - \alpha$. Les bornes de cet intrevalle dépendent donc de θ. Soient $b_1(\theta)$ et $b_2(\theta)$ les bornes de cet intervalle.

$$1 - \alpha = P[b_1(\theta) < \hat{\theta} < b_2(\theta)]$$

On va ensuite réécrire cet intevalle pour obetnir la forme suivante :

$$1 - \alpha = P[a_1(\hat{\theta}) < \theta < a_2(\hat{\theta})]$$

Les bornes dépendent cette fois de $\hat{\theta}$.

Intervalle de confiance pour le loyer

On choisit ici $\alpha = 0,05$

Soit $\bar{y}$ notre estimateur du loyer moyen m.

$$1 - \alpha = P[\bar{y} - a < m < \bar{y} + a]$$

$$\iff 1-\alpha = P[-a < \bar{y} - m < a]$$

On va maintenant utiliser l'estimation de l'écart type de $\bar{y}$ pour réduire la variable aléatoire :

$$\iff 1-\alpha = P[\frac{-a}{\sigma/\sqrt{n}} < \frac{\bar{y}-m}{\sigma/\sqrt{n}} < \frac{a}{\sigma/\sqrt{n}}]$$

Avec $\sigma/\sqrt{n}$ l'écart-type de l'estimation du loyer moyen.

Comme on sait que le loyer moyen suit une loi Normale si n est suffisamment grand, on sait que la variable $\frac{\bar{y}-m}{\sigma/\sqrt{n}}$ suit une loi Normale centrée réduite.

Soit $F(.)$ la fonction de répartition de la loi Normale centrée réduite.

$$\iff 1-\alpha = F(\frac{a}{\sigma/\sqrt{n}}) - F(\frac{-a}{\sigma/\sqrt{n}}) = 2F(\frac{a}{\sigma/\sqrt{n}}) - 1$$

Connaisant α, n et σ(ou au moins un estimateur de σ), on va chrecher a qui satisfasse l'expression précédente.

Dans le TD précédent, on avait pour le premier échantillon $n_1 = 6$, $\bar{y} = 493$ et $\sigma_{\bar{y}} = \sqrt{5844,44} \approx 76$

$$0,95 = 2 * S_{(5)}(\frac{a}{76}) - 1 \iff S_{(5)}(\frac{a}{76}) = \frac{1,95}{2} = 0,975$$

Ici, n est trop petit pour utiliser l'approximation normale, on utilise donc la loi de Student à $n-1$ degrés de liberté. Soit $S_k(.)$la fonction de répartion de la loi de Student à k degrés de liberté.

En cherchant 0,975 dans la table de la loi de Student à 5 degrés de liberté, on en déuit :

$$\frac{a}{76} = 2,571 \iff a = 2,571 * 76 = 195,4$$

D'où l'intervalle suivant obtenu à partir de l'echantillon 1 :

$$0,95 = P\left[297,6 < m < 688,4\right]$$

Ce qui signifie que la moyenne des loyers est compris avec 95% de certitude dans cet intervalle.

Pour le second échantillon, on avait $n_2 = 300$, $\bar{y} = 550$ et $\sigma_{\bar{y}} = \sqrt{300} \approx 17$

On peut cette fois utiliser l'approximation normale.

$$F(\frac{a}{17}) = 0,975$$

De la même façon, en lisant dans la table de la loi Normale centrée réduite, on en déduit :

$$\frac{a}{17} = 1,96 \iff a = 1,96 * 17 = 33,32$$

D'où l'intervalle suivant obtenu à partir de l'échantillon 2 :

$$0,95 = P\left[516,68 < m < 583,32\right]$$

Ce résultat illustre le fait qu'en prenant davantage d'observations on a accru la précision de l'estimation. Plus le nombre d'obsrevations est important, plus on se rapproche de la vraie valeur du paramètre estimé.

Question 7 - Intervalle de confiance (bis)

Si l'on considère que l'on connaît réellement la variance, alors le raisonnement ne change pas pour le second échantillon, par contre on sait maintenant que l'estimateur du loyer moyen calculé à partir du premier échantillon suit bien une loi Normale. On en déduit :

$$F(\frac{a}{76}) = \frac{1,95}{2} = 0,975$$

D'où :

$$\frac{a}{76} = 1,96 \Longleftrightarrow a = 1,96 * 76 = 148,96$$

Il vient alors :

$$0,95 = P\left[344,04 < m < 641,96\right]$$

En supposant qu'on connaisse la variance de l'échantillon, on a réduit l'intervalle de confiance, la précision est donc plus importante pour un échantillon de petite taille. En revanche, pour le second échantillon, de taille plus importante, il n'y a aucune modification. Ceci vient du fait que l'estimation est convergente : pour un grand échantillon, il est équivalent de disposer de la vraie valeur ou d'une estimation de cette valeur.

Exercice 2

Question 1 - Modèle de prévision

Envisageons le modèle suivant :

$$y_i = b_0 + b_1 S_i + b_2 A_i + b_3 R_i + b_4 T_i + b_5 C_i + \varepsilon_i$$

Avec :

- y_i le loyer du studio
- S_i la surface du studio en m^2
- A_i l'ancienneté de l'immeuble en nombre d'années
- R_i une variable qui prend la valeur 1 si le quartier est résidentiel, 0 sinon
- T_i une variable qui prend la valeur 1 si les transports en commun sont proches, 0 sinon
- C_i une variable qui prend la valeur 1 si les commerces sont proches, 0 sinon

Remarquez qu'il faut définir la notion de proximité. On pourrait aussi choisir une mesure de distance, dans ce cas les variables T_i et C_i ne serait plus des variables indicatrices, mais des distances.

La variable ε_i représente l'influence sur le loyer d'un studio de toutes les variables non prises en compte par le modèle. On suppose que cette variable aléatoire est d'espérance nulle.

Ce modèle fait appel à la théorie des prix hédoniques : le prix d'un bien est déterminé par l'addition des valeurs de l'ensemble de ses caratèristiques.

Question 2 - Perturbations

Dans ce modèle, les facteurs omis qui influencent le prix peuvent être par exemple l'équipement intérieur, l'existence d'un ascenseur, l'étage, la proximité d'un espace vert, la mise à disposition d'une cave ou d'un garage ou l'exposition au soleil

Question 3 - Moindres carrés et maximum de vraisemblance

Analysons les hypothèses de l'énoncé :

Les perturbations sont distribuées selon une loi Normale : $\varepsilon_i \rightarrow N(0;\sigma^2)$

Les variables explicatives ne sont pas aléatoires : $E(b_0 + b_1S_i + b_2A_i + b_3R_i + b_4T_i + b_5C_i) = b_0 + b_1S_i + b_2A_i + b_3R_i + b_4T_i + b_5C_i$

On peut en déduire que y_i suit une loi Normale de variance σ^2 et d'espérance m, avec :

$$m = E(y_i) = E(b_0+b_1S_i+b_2A_i+b_3R_i+b_4T_i+b_5C_i+\varepsilon_i) = b_0+b_1S_i+b_2A_i+b_3R_i+b_4T_i+b_5C_i$$

Ecrivons la fonction de log-vraisemblance de l'échantillon :

$$ln[L(.)] = -\frac{n}{2}ln(2\Pi) - \frac{n}{2}ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - b_0 - b_1S_i - b_2A_i - b_3R_i - b_4T_i - b_5C_i)^2$$

Pour maximiser, on dérive par rapport à tous les paramètres d'intérêt :

$$\frac{\partial ln[L(.)]}{\partial b_0} = -\frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - b_0 - b_1S_i - b_2A_i - b_3R_i - b_4T_i - b_5C_i) = 0$$

$$\frac{\partial ln[L(.)]}{\partial b_1} = -\frac{1}{\sigma^2}\sum_{i=1}^{n}S_i(y_i - b_0 - b_1S_i - b_2A_i - b_3R_i - b_4T_i - b_5C_i) = 0$$

$$\frac{\partial ln[L(.)]}{\partial b_2} = -\frac{1}{\sigma^2}\sum_{i=1}^{n}A_i(y_i - b_0 - b_1S_i - b_2A_i - b_3R_i - b_4T_i - b_5C_i) = 0$$

$$\frac{\partial ln[L(.)]}{\partial b_3} = -\frac{1}{\sigma^2}\sum_{i=1}^{n}R_i(y_i - b_0 - b_1S_i - b_2A_i - b_3R_i - b_4T_i - b_5C_i) = 0$$

$$\frac{\partial ln[L(.)]}{\partial b_4} = -\frac{1}{\sigma^2}\sum_{i=1}^{n}T_i(y_i - b_0 - b_1S_i - b_2A_i - b_3R_i - b_4T_i - b_5C_i) = 0$$

$$\frac{\partial ln[L(.)]}{\partial b_5} = -\frac{1}{\sigma^2}\sum_{i=1}^{n}C_i(y_i - b_0 - b_1S_i - b_2A_i - b_3R_i - b_4T_i - b_5C_i) = 0$$

Ecrivons la fonction de perte utilisée par la méthode des moindres carrés :

$$S = \sum_{i=1}^{n}(y_i - b_0 - b_1S_i - b_2A_i - b_3R_i - b_4T_i - b_5C_i)^2$$

Pour minimiser, on dérive par rapport à tous les paramètres d'intérêt :

$$\frac{\partial S}{\partial b_0} = -2\sum_{i=1}^{n}(y_i - b_0 - b_1 S_i - b_2 A_i - b_3 R_i - b_4 T_i - b_5 C_i) = 0$$

$$\frac{\partial S}{\partial b_1} = -2\sum_{i=1}^{n} S_i(y_i - b_0 - b_1 S_i - b_2 A_i - b_3 R_i - b_4 T_i - b_5 C_i) = 0$$

$$\frac{\partial S}{\partial b_2} = -2\sum_{i=1}^{n} A_i(y_i - b_0 - b_1 S_i - b_2 A_i - b_3 R_i - b_4 T_i - b_5 C_i) = 0$$

$$\frac{\partial S}{\partial b_3} = -2\sum_{i=1}^{n} R_i(y_i - b_0 - b_1 S_i - b_2 A_i - b_3 R_i - b_4 T_i - b_5 C_i) = 0$$

$$\frac{\partial S}{\partial b_4} = -2\sum_{i=1}^{n} T_i(y_i - b_0 - b_1 S_i - b_2 A_i - b_3 R_i - b_4 T_i - b_5 C_i) = 0$$

$$\frac{\partial S}{\partial b_5} = -2\sum_{i=1}^{n} C_i(y_i - b_0 - b_1 S_i - b_2 A_i - b_3 R_i - b_4 T_i - b_5 C_i) = 0$$

Ce sont donc bien les même équations qui vont conduire à l'obtention des estimateurs par les deux méthodes.

Exercice 3

Question 1 - Méthode des moments

L'estimateur de p par la méthode des moments est la fréquence empirique.

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Où x_i est une variable indicatrice qui prend la valeur 1 si l'individu a trouvé un CDI, 0 sinon.

Question 2 - Méthode du maximum de vraisemblance

La variable x_i suit une loi de Bernouilli. On sait donc que $P(x_i = 1) = p$ et $P(x_i = 0) = 1 - p$

Ecrivons la densité de la variable : $f(x_i) = p^{x_i}(1-p)^{x_i}$

Sous l'hypothèse d'indépendance des observations, on peut écrire la fonction de log-vraisemblance :

$$ln[L(.)] = \sum_{i=1}^{n} x_i ln(p) + (1 - x_i) ln(1 - p)$$

Annulons sa dérivée :

$$\frac{\partial ln[L(.)]}{\partial p} = 0 \Longleftrightarrow \frac{1}{p}\sum_{i=1}^{n} x_i - \frac{1}{1-p}\sum_{i=1}^{n}(1-x_i) = 0$$

$$\frac{1-p}{p(1-p)}\sum_{i=1}^{n} x_i - \frac{p}{p(1-p)}\sum_{i=1}^{n}(1-x_i) = 0 \Longleftrightarrow (1-p)\sum_{i=1}^{n} x_i - p\sum_{i=1}^{n}(1-x_i) = 0$$

$$\Longleftrightarrow (1-p)\frac{1}{n}\sum_{i=1}^{n} x_i - p\frac{1}{n}\sum_{i=1}^{n}(1-x_i) = 0 \Longleftrightarrow (1-p)\bar{x} - p\sum_{i=1}^{n}\frac{1}{n} + p\bar{x}) = 0$$

Avec $\bar{x}$ la moyenne empirique de la distribution.

$$\Longleftrightarrow (1-p)\bar{x} - p + p\bar{x} = 0 \Longleftrightarrow \bar{x} - p\bar{x} - p + p\bar{x} = 0$$

$$\hat{p} = \bar{x}$$

On retrouve ici le résultat précédent.

Question 3 - Application numérique

Estimation de p : $\hat{p} = 235/300 = 0,783$

Question 4 - Intervalle de confiance

La variance de notre estimation est $\sigma_p^2 = \frac{p(1-p)}{n}$
Notre intervalle de confiance pour p est le suivant :

$$0,95 = P\left[0,783 - 1,96\sqrt{\frac{p(1-p)}{n}} < p < 0,783 + 1,96\sqrt{\frac{p(1-p)}{n}}\right]$$

Si l'on estime la variance de p à l'aide de $\hat{p}$, on obtient l'intervalle à 95% suivant :

$$[0,736; 0,830]$$

Question 5 - Intervalle de confiance (bis)

On réduit l'échantillon à 20 observations. En conservant l'approximation précédente, on obtient l'intervalle suivant :

$$[0,602; 0,964]$$

Exercice 4

Question 1 - Modèlisation

On peut choisir une loi de Poisson dont la densité s'écrit $P(Y = k) = exp(-\lambda)\frac{\lambda^k}{k!}$

Question 2 - Estimateur du maximum de vraisemblance

Suposons qu'on dispose de n observations. La variable x_i est le nombre de lettres que l'individu i a envoyé pour trouver un stage ou un emploi.

La fonction de vraisemblance de notre échantillon s'écrit :

$$L(.) = \prod_{i=1}^{n} exp(-\lambda)\frac{\lambda^{x_i}}{x_i!} = exp(-n\lambda)\frac{\lambda^{(x_1+x_2+...+x_n)}}{x_1!x_2!...x_n!}$$

On peut en déduire la fonction de log-vraisemblance :

$$ln[L(.)] = -n\lambda + ln(\lambda)\sum_{i=1}^{n} x_i - ln(x_1!x_2!...x_n!)$$

Annulons la dérivée par rapport à λ :

$$\frac{\partial ln[L(.)]}{\partial \lambda} = -n + \frac{1}{\lambda}\sum_{i=1}^{n} x_i = 0$$

$$\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Question 3 - Intervalle de confiance

On suppose $\hat{\lambda} = 6$

On désire construire un intervalle tel que $P(\lambda_{inf} < \lambda < \lambda_{sup}) = 0,95$

Choisissons les bornes telles que : $\lambda_{inf} = \hat{\lambda} - a$ et $\lambda_{sup} = \hat{\lambda} + a$

Il vient alors : $P(-a < \hat{\lambda} - \lambda < a) = 0,95$

On va estimer l'écart-type par $\sqrt{\hat{\lambda}/n} = \sqrt{6/20} = 0,547$

On peut alors écrire : $P(-\frac{a}{0,547} < \frac{\hat{\lambda}-\lambda}{0,547} < \frac{a}{0,547}) = 0,95$

Si on suppose que n est suffisamment grand, on peut dire que $\frac{\hat{\lambda}-\lambda}{0,547}$ suit une loi Normale centrée réduite.

On utilise la table de la loi Normale pour trouver $a = 1,96 * 0,547$

D'où l'intervalle de confiance à 95% pour la vraie valeur de λ :

$$[6 - 1,96 * 0,547; 6 + 1,96 * 0,547] \Longleftrightarrow [4,928; 7,072]$$

Question 4 - Intervalle de confiance (bis)

Intervalle de confiance à 90%

On veut cette fois :

$$P(-\frac{a}{0,547} < \frac{\hat{\lambda} - \lambda}{0,547} < \frac{a}{0,547}) = 0,90$$

$$\Longleftrightarrow F(\frac{a}{0,547}) - F(-\frac{a}{0,547}) = 0,90 \Longleftrightarrow 2F(\frac{a}{0,547}) - 1 = 0,90$$

Avec F(.) la fonction de répartition de la loi Normale centrée réduite.

$$F(\frac{a}{0,547}) = \frac{1,90}{2} = 0,95$$

On en déduit :

$$\frac{a}{0,547} = 1,645 \Longleftrightarrow a = 0,8998$$

D'où l'intervalle de confiance à 90% suivant :

$$[5,1002; 6,8998]$$

Intervalle de confiance à 99%

De la même façon que précédemment, on veut cette fois : $F(\frac{a}{0,547}) = 0,995$
On en déduit :

$$\frac{a}{0,547} = 2,575 \Longleftrightarrow a = 1,408$$

D'où l'intervalle de confiance à 90% suivant :

$$[4,592; 7,408]$$

Remarque

On observe que l'intrevalle grandit avec le degré de confiance exigé.

Question 5 - Accroissement de l'échantillon

Accroître la taille de l'échantillon permettrait de renforcer la précision de l'estimation et de confirmer la possibilité d'approximer par la loi Normale.

Question 6 - Projection

Le nombre de lettre que vous pensez envoyer dépend de plusieurs paramètres, notamment votre aversion pour le risque, votre jugement sur la qualité de votre propre CV, le caractère ciblé ou non de vos démarches, etc...