Statistiques Appliquées
Contrôle Continu N°2 - Eléments de correction
TD N°8 et 9

Mathieu Valdenaire

Partie A : Dépouillement et pronostic sur les résultats d'une élection (5,5 points)

Dans une élection à un tour, 20000 votants sont appelés à se prononcer pour un des deux candidats, John et Hilary.
Les premiers dépouillements indiquent que sur 1% des votants (choisis de manière aléatoire), John recueille 45% des suffrages.

1. **Cela signifie-t-il qu'Hilary va l'emporter ? Définir l'outil que vous utilisez pour répondre à cette question et y apporter une réponse, sachant que l'on veut pouvoir accorder** 98% **de confiance à cette réponse.**

   On veut savoir si ce pourcentage de 45% des voix obtenu sur une petite partie de la population permet d'affirmer, avec 98% de confiance, que le pourcentage obtenu sur l'ensemble de la population est inférieur à 50%. Il s'agissait donc ici de construire une borne supérieure de confiance de niveau 98% sur le pourcentage des voix obtenu par John, p (ceux qui ont calculé une borne inférieure de confiance sur le score d'Hilary ont obtenu la grande majorité des points - il s'agissait en tout cas de calculer une borne, supérieure ou inférieure selon le candidat auquel on s'intéresse, et non un intervalle de confiance bilatéral comme beaucoup l'ont fait).
   Soit X l'indicatrice de "favorable à John", et p la proportion d'électeurs favorables à John parmi les 20.000 électeurs. Le nombre n d'électeurs interrogés étant très petit devant la taille de la population, on peut considérer que les n variables tirées sont indépendantes et identiquement distribuées, comme si le tirage avait été fait "avec remise". Le modèle que nous utilisons est donc : $X_1, ..., X_n \approx i.i.d.B(1, p)$ (n-échantillon d'une loi de Bernoulli de paramètre p).
   Pour construire un intervalle de confiance, nous devons utiliser une "fonction pivotale", c'est-à-dire une variable fonction uniquement des observations et du paramètre étudié, dont la loi soit entièrement connue. La fonction pivotale utilisée ici est construite à partir de la proportion observée d'électeurs favorables à John parmi les n interrogés. Nous considérons que n est assez grand pour pouvoir utiliser l'approximation normale de la loi de cette proportion, notée F_n :

$$F_n = \frac{1}{n}\sum_{i=1}^{n} X_i \approx N\left(p, \frac{p(1-p)}{n}\right) \Leftrightarrow U_n = \frac{F_n - p}{\sqrt{\frac{F_n(1-F_n)}{n}}} \approx N(0;1)$$

   Lecture dans la table de la loi $N[0;1]$: $P\{U \leq 2,056\} = 0,98$

   Soit :

$$P\{-2,056 \leq U\} = 0,98$$

$$P\left\{-2,056 \leq \frac{F_n - p}{\sqrt{\frac{F_n(1-F_n)}{n}}}\right\} \simeq 0,98$$

$$P\left\{-F_n - 2,056\sqrt{\frac{F_n(1-F_n)}{n}} \leq -p\right\} \simeq 0,98$$

$$P\left\{F_n + 2,056\sqrt{\frac{F_n(1-F_n)}{n}} \geq p\right\} \simeq 0,98$$

$$P\left\{p \leq F_n + 2,056\sqrt{\frac{F_n(1-F_n)}{n}}\right\} \simeq 0,98$$

La borne supérieure de confiance à 98% est donc égale à $F_n + 2,056\sqrt{\frac{F_n(1-F_n)}{n}}$.
La proportion observée de $f_n = 0,45$ conduit à $p \leq 0,45 + 0.072$, soit $p \leq 0,522$.
Rappel : il n'est pas question d'écrire la probabilité que p soit supérieur à 0,522, car il n'y a plus rien d'aléatoire dans l'inégalité ; une fois l'observation faite, la proposition est soit vraie (proba 1) soit fausse (proba 0).

2. **A partir de quel niveau de confiance pourra-t-on se dire certain de la défaite de John, dans ces conditions ?**
   L'expression d'une borne supérieure de confiance $A(F_n)$ de niveau $1-\alpha$ était donc $F_n + u_{1-\alpha}\sqrt{\frac{F_n(1-F_n)}{n}}$.
   Si on veut être certain de la défaite de John avec un degré de confiance $1-\alpha$, il nous faut avoir :

$$F_n + u_{1-\alpha}\sqrt{\frac{F_n(1-F_n)}{n}} \leq 0,50$$

   Cette condition aboutit à :

$$u_{1-\alpha} \leq \frac{0,50 - F_n}{\sqrt{\frac{F_n(1-F_n)}{n}}}$$

   Soit avec $F_n = 0,45$:

$$u_{1-\alpha} \leq 1,42$$

   Cette borne de 1,42 correspond (lecture dans la table de la loi Normale) à un niveau de confiance de $92,22\%$. L'observation, sur une sous-population de 200 individus tirés de manière aléatoire, d'une proportion d'électeurs favorables à John de 45% permet donc d'affirmer que sa défaite est certaine avec $92,2\%$ de confiance au plus.

Partie B : Nombre d'essais pour remporter une élection (5,5 points)

Le candidat malheureux à l'élection précédente se demande désormais combien de fois il va devoir se présenter avant de remporter une élection.
Ce type de phénomène (nombre d'essais pour faire apparaître un événement de probabilité p), est modélisé par une loi géométrique (on fait l'hypothèse que la probabilité de succès est la même à chaque élection).
La fonction de probabilité de la variable X (nombre d'essais pour remporter une élection), suivant une loi géométrique de paramètre p est la suivante :

$$Pr[X_i = x] = p(1-p)^{x-1}, \quad avec \quad x = 1, 2, ...$$

Il vous est demandé d'estimer le paramètre de cette loi. On peut pour cela se fonder sur l'observation du temps d'attente de ses prédécesseurs, en supposant que la variable X est distribuée de manière identique. Cette observation, pour ses douze prédécesseurs, aboutit à l'échantillon :

1 1 6 4 1 2 3 2 1 4 2 1

1. **Donner une estimation du paramètre de cette loi par la méthode des moments. L'espérance d'une loi géométrique est $E(X) = 1/p$ et sa variance $V(X) = (1-p)/p^2$).**

   Le moment théorique d'ordre 1, rappelé dans l'énoncé, nous donne immédiatement $p = 1/E(X)$. On passe aux moment empirique en appliquant cette formule à notre échantillon, soit : $\widehat{p} = 1/\overline{X} = \frac{1}{\sum x_i/n} = \frac{n}{\sum x_i} = 12/28$.

2. **Calculer l'estimateur du maximum de vraisemblance de ce paramètre et en déduire une estimation du paramètre dans le cas présent. Commenter.**

   La vraisemblance de l'échantillon s'écrit :
   $l(p; x_1, x_2, ..., x_n) = \prod_{i=1}^{n} f_p(x_i)$
   Soit, ici :
   $l(p; x_1, x_2, ..., x_n) = \prod_{i=1}^{n} p(1-p)^{x_i - 1} = p^n (1-p)^{\sum x_i - n}$

   Rappel : la probabilité d'apparition a priori d'un échantillon est égale au produit des probabilités d'apparition de chacune des réalisations de cet échantillon (puisque celles-ci sont supposées indépendantes deux à deux). C'est ce produit qui est appelé fonction de vraisemblance. La méthode du maximum de vraisemblance consiste ensuite à rechercher la valeur du paramètre qui rend cette probabilité maximale, afin de déterminer la valeur numérique la plus vraisemblable pour le paramètre, étant donnée l'observation de l'échantillon obtenu.

   La log-vraisemblance de l'échantillon s'écrit :
   $$\begin{aligned} L(p; x_1, x_2, ..., x_N) &= ln(l(p; x_1, x_2, ..., x_N)) \\ &= ln(p^n (1-p)^{\sum x_i - n}) \\ &= n ln(p) + (\sum x_i - n) ln(1-p) \end{aligned}$$

   Rappel 2 : le passage au logarithme sur la fonction permet de passer de la maximisation d'un produit à celle d'une somme, ce qui facilite le calcul, le résultat restant le même car la fonction logarithme est monotone strictement croissante.
   L'estimateur du maximum de vraisemblance réalise le maximum en p de L(p).

   La condition nécessaire d'ordre 1 (CN1) est $\frac{\partial L(p)}{\partial p} = 0$ en $p = \widehat{p}$, soit ici :
   $n * \frac{1}{\widehat{p}} + (\sum x_i - n) * \frac{-1}{1-\widehat{p}} = 0$
   soit après simplification $\widehat{p} = \frac{n}{\sum x_i} = \frac{1}{\overline{X}}$
   Remarque : une condition suffisante d'ordre 2 (CS2) est $\frac{\partial^2 L(p)}{\partial p^2} < 0$ en $p = \widehat{p}$

Partie C : Prévision du nombre de voix à la prochaine élection (4 points)

Le candidat malheureux à la première élection souhaite désormais prévoir le nombre de voix qu'il obtiendra à la prochaine élection (variable Y).

On retient désormais l'approximation de la loi de la variable Y par la loi normale pour la suite de l'exercice. On estime les paramètres de la loi de cette variable Y à partir de l'observation du score des candidats du même parti aux précédentes élections (70 observations), en supposant qu'à chaque élection le nombre de leurs voix est indépendant et identiquement distribué. La proportion d'électeurs votant pour le candidat en question est ainsi estimée à 48% des voix, soit 9600 voix, et la variance estimée est égale à 4992.

Entre quelles bornes le pourcentage de voix de ce candidat à la prochaine élection se situera-t-il ? (à 95% de confiance) ?

On souhaite *prévoir* (c'est donc un intervalle de *prévision*) le *nombre* de voix du candidat (et non la *proportion* des voix obtenue) : il faut donc modéliser le nombre de voix Y, et non une proportion F_n.

On a supposé que le nombre de voix à la prochaine élection Y_{n+1} est indépendant et identiquement distribué au nombre de voix obtenu par le candidat du même parti aux élections précédentes. La meilleure prévision de Y_{n+1} est $\overline{Y}$.

Cette variable $\overline{Y}$ suit une loi de Student à 70-1 degrés de liberté, que l'on peut donc approximer par une loi normale, cette approximation étant considérée comme valable au-delà de 30 degrés de liberté.

L'erreur de prévision $Y_{n+1} - \overline{Y}$ suit une loi $N\left(0; 4992(1+\frac{1}{70})\right)$

(car $E(Y_{n+1} - \overline{Y}) = E(Y_{n+1}) - E(\overline{Y}) = m - m = 0$

et $V(Y_{n+1} - \overline{Y}) = V(Y_{n+1}) + V(\overline{Y}) - 2cov(Y_{n+1}; \overline{Y}) = \sigma^2 + \frac{\sigma^2}{N} = \sigma^2(1+\frac{1}{N})$.

La fonction pivotale que l'on utilise est la suivante :

$$U = \frac{\left(Y_{n+1} - \overline{Y}\right) - \left(E(Y_{n+1}) - E(\overline{Y})\right)}{\sqrt{V(Y_{n+1}) + V(\overline{Y})}} = \frac{\left(Y_{n+1} - \overline{Y}\right) - 0}{\sqrt{\sigma^2(1+\frac{1}{N})}} \approx N(0;1)$$

Un intervalle bilatéral à 95% correspond à une probabilité d'erreur de 2,5% de chaque côté. Lecture dans la table de la loi $N(0;1)$ pour un intervalle bilatéral de confiance à 99% :

$P\{U \leq 1,96\} = 0,975$:

Ainsi : $P\{-1,96 \leq U \leq 1,96\} = 0,95$

Résolvons les inégalités en Y_{n+1} , afin d'obtenir l'intervalle bilatéral de prévision à 95% pour Y_{n+1} :

$$P\left(-1,96 \leq \frac{\left(Y_{n+1} - \overline{Y}\right)}{\sqrt{\sigma^2\left(1+\frac{1}{N}\right)}} \leq 1,96\right) = 0,99$$

$$P\left(-1,96\sqrt{\sigma^2\left(1+\frac{1}{N}\right)} \leq Y_{n+1} - \overline{Y} \leq 1,96\sqrt{\sigma^2\left(1+\frac{1}{N}\right)}\right) = 0,95$$

$$P\left(\overline{Y} - 1,96\sqrt{\sigma^2\left(1+\frac{1}{N}\right)} \leq Y_{n+1} \leq \overline{Y} + 1,96\sqrt{\sigma^2\left(1+\frac{1}{N}\right)}\right) = 0,95$$

L'application numérique l'intervalle de prévision à 95% pour le nombre de voix à la prochaine élection est donc :

$$9600 - 1,96\sqrt{4992\left(1+\frac{1}{71}\right)} \leq Y_{n+1} \leq 9600 + 1,96\sqrt{4992\left(1+\frac{1}{71}\right)}$$

$$9600 - 139,47 \leq Y_{n+1} \leq 9600 + 139,47$$

$$9460 \leq Y_{n+1} \leq 9740$$

Remarque : si l'intervalle de prévision avait porté sur la proportion des voix obtenue F_n, l'intervalle aurait été $47,3\%$ - $48,7\%$. Ceux qui sont parvenus à ce résultat ne se sont vus retirer qu'une partie marginale des points.

Partie D : Modélisation économétrique du nombre de voix aux élections (5 points)

Les modèles les plus performants de prévision électorale font dépendre le nombre de voix de la majorité sortante au second tour d'une élection des variables suivantes[1] :

$$VOT_{it} = c_i + \alpha_1 DPIB_t + \alpha_2 CHO_{it} + \alpha_3 POP_t + \alpha_4 PRE_{it} + \alpha_5 VP_{it} + \alpha_6 ELI_{it} + \epsilon_{it}$$

où la variable dépendante VOT représente le nombre de voix obtenu à l'élection, et les variables explicatives retenues sont $DPIB_t$, la différence entre le taux de croissance réelle du PIB l'année des élections et celui de l'année précédant l'année des élections, POP_t la popularité du Premier ministre, PRE_{it} est une variable politique constituée par les résultats des élections précédentes, VP_{it} une variable reflétant les différences partisanes entre le département où a lieu le vote et la moyenne au niveau national, et enfin une variable ELI_{it} qui reflète la probabilité pour la majorité sortante d'avoir été éliminée dès le premier tour.

1. **Quel est l'intérêt d'un tel modèle pour prévoir le résultat des élections ?** (1 point)

   Le but de cette question était de comparer l'intérêt de prévoir les résultats d'une élection à partir d'enquêtes d'opinion (sondages) en formant des intervalles de confiance sur la proportion de personnes se déclarant prêtes à voter pour tel ou tel candidat et la prévision économétrique réalisée à partir d'un tel modèle. (toutes sortes de réponses ont été acceptées). Un premier intérêt est de pouvoir réaliser une estimation plus précise car cette méthode permet d'avoir recours à des échantillons plus grands, les variables utilisées étant disponibles pour un grand nombre d'individus (par rapport à un sondage, dont le coût limite souvent la taille de l'échantillon - en pratique, les sondages électoraux sont souvent réalisés sur des échantillons d'environ 1000 individus). Un second intérêt est de s'affranchir des biais éventuels des sondages, liés entre autres à la subjectivité des réponses, à la mise en situation des répondants, à la formulation de la question... De fait, ce modèle est très performant : son estimation sur les données des élections législatives de 2002 a permis de prévoir les résultats du second tour à 4 sièges près.

2. **Quelle est la signification du coefficient α_1 ? Quel signe peut-on attendre pour ce coefficient ? (1 point)**

   Le coefficient reflète l'effet sur le nombre de voix obtenu à l'élection de l'augmentation d'une unité de la variable DPIB, autrement dit de l'augmentation d'un point de la croissance entre l'année de l'élection et l'année précédente. Etant donné qu'une augmentation de DPIB reflète une amélioration de la situation

[1] Le modèle présenté ici est inspiré de Dubois et Auberger (2003), "Situation politico-économique et résultats des élections législatives françaises", *Revue Economique* 54(3), pp 551-560.

économique on peut s'attendre à un effet positif, la variable à laquelle on s'intéresse étant le nombre de voix de la majorité sortante, à laquelle cette amélioration devrait a priori bénéficier électoralement.
L'estimation de ce modèle par la méthode des moindres carrés ordinaires (MCO) sur les données des élections législatives sur la période 1981-1997 donne les résultats suivants :

$$VOT_{it} = c_i + 0.91DPIB_t + 4.76POP_t + 0.54PRE_{it} + 0.30VP_{it} - 0.29ELI_{it} + e_{it}$$

$$(0.210) \qquad (0.889) \qquad (0.040) \qquad (0.037) \qquad (0.012)$$

(les chiffres entre parenthèses sont les écart-types estimés des coefficients correspondants de l'équation ci-dessus).

3. **Pourquoi le coefficient d'une régression de MCO est-il une variable aléatoire ?** (0,5 point)

   Dans le modèle linéaire, les variables dépendantes sont des variables aléatoires car elles sont généralement obtenues par échantillonnage aléatoire dans une population. Les estimateurs des MCO sont des fonctions de ces variables, et par conséquent sont également des variables aléatoires (Cf TD3)

4. **Rappeler les propriétés des estimateurs des MCO** (1 point)

   Sous les hypothèses habituelles, l'estimateur des moindres carrées ordinaires (qui consiste à rechercher les valeurs des paramètres qui minimisent la somme des carrés des erreurs) est linéaire, sans biais ($E(\hat{a}) = a$) et efficace (de variance minimale) i.e. BLUE (Best Linear Unbiased Estimator) c'est à dire qu'il n'existe pas d'estimateur sans biais de a, qui ait une variance plus petite.

5. **La popularité du premier ministre, reflétée par la variable POP_t a-t-elle un impact significatif sur le nombre de voix obtenu par la majorité sortante aux élections ?**(1.5 point)

   Cette variable a un impact significatif si le coefficient qui lui est affecté peut être considéré comme significativement différent de zéro, autrement dit si l'intervalle de confiance qui peut être formé autour du coefficient ne contient pas 0. Au seuil de 5% de confiance l'intervalle de confiance estimé autour du coefficient de la variable POP est $[4,76 - 1,96 * 0,889; 4,76 + 1,96 * 0,889]$, soit [3,02 ; 6,50], on peut donc considérer, au seuil de 5%, que cette variable a un effet significatif (et positif) sur le nombre de voix obtenu (tous ceux qui ont écrit quelque chose comme "oui car 4.76>2*0.889" ont obtenu les points pour cette question).