MSI : éléments de correction du TD3
Intervalles de confiance

ENTPE 2A

5 Février 2015

Préambule : rappel du cours en amphi

Exercice 1 : durées de vie

Nous souhaitons estimer les paramètres m et σ^2 de l'ensemble de la population des piles, c'est-à-dire l'espérance de la durée de vie d'une pile et sa variance. Nous tirons un échantillon de 10 piles et nous pouvons calculer des statistiques sur cet échantillon, la moyenne empirique $\bar{x}$ et la variance empirique S^2, qui vont servir à donner une estimation de m et σ^2. Attention à ne pas confondre les variables à estimer et les caractéristiques de l'échantillon. Faire de l'inférence statistique, en particulier de l'estimation, c'est essayer de donner une estimation d'un ou plusieurs paramètres d'une population à partir d'un ensemble restreint de données, l'échantillon. Ici, l'hypothèse est que la population (les piles auxquelles on associe leur durée de vie) se distribue selon une loi normale $N(m, \sigma)$ dont on veut estimer les paramètres.

La première partie consiste à estimer m avec σ^2 inconnue. Il s'agit du cas 1 vu en cours. On va encadrer l'estimation de m par deux bornes construites à partir de caractéristiques connues (voir cours). Pour cela, on cherche une fonction pivotale, c'est-à-dire une combinaison des variables connues, qui va suivre une loi dont on connait la fonction de répartition. Dans vos tables, vous avez trois lois disponibles : la loi normale centrée-réduite, la loi de Student (qui est symétrique, centrée en 0 comme la loi normale $N(0,1)$ et la loi du χ^2 (qui n'est pas symétrique et définie sur $\Re^+$. Comme σ^2 est inconnue, on prend son estimateur s^2 vu en cours et la fonction pivotale est

$$T = \frac{\bar{X} - m}{S}\sqrt{n-1}$$

qui suit une loi de Student à (n-1) degrés de libertés. L'intervalle de confiance est donné par la formule

$$\bar{x} - t\frac{s}{\sqrt{n-1}} < m < \bar{x} - t\frac{s}{\sqrt{n-1}}$$

où t est le fractile d'ordre $1 - \frac{\alpha}{2}$ de la loi de Student. Votre table donne les fractile d'ordre P pour certains degrés de liberté. Il faut donc lire ici la valeur pour $\nu = 9$, soit $10 - 1$ et $P = 0.975$. On trouve $t = 2.262$. L'application numérique donne l'intervalle suivant pour m :

$$m \in \left[754, 764\right]$$

Pour l'estimation de σ^2, on se place également dans le cas m inconnue. Les deux questions auraient pu être faites dans l'ordre inverse. Comme m est inconnue, nous sommes dans le cas numéro 4 vu en cours. La fonction pivotale utilisée est

$$\frac{nS^2}{\sigma^2}$$

qui suit une loi $\chi^2 n - 1$. L'intervalle de confiance est

$$\frac{ns^2}{l_2} < \sigma^2 < \frac{ns^2}{l_1}$$

l_1 et l_2 sont les fractiles d'ordre $\frac{\alpha}{2}$ et $1 - \frac{\alpha}{2}$ de la loi du $\chi^2 n - 1$. Pour un niveau de risque de 5%, pour $n - 1 = 9$ degrés de libertés, on lit dans la table $l_1 = 2.70$ et $l_2 = 19.02$. Ensuite il suffit de faire une application numérique et on trouve quelque chose comme $\sigma^2 \in [19.9, 140.2]$. Cet intervalle de confiance vous a paru grand en TD mais il faut savoir qu'il s'agit d'une variance, donc le carré de l'écart type, ceci n'est pas aberrant.

Exercice 2 : albinos

On observe une population animale dont certains membres sont albinos. On a extrait un échantillon de 400 animaux dont 30 sont albinos. Construire un intervalle de confiance à 95% pour la population d'albinos.
Il faut bien comprendre que nous nous intéressons à la proportion d'albinos et que l'on cherche à estimer dans quel intervalle serait cette proportion pour la population totale. On note f la proportion trouvée dans l'échantillon, et p la proportion dans la population. Ici, $f = \frac{30}{400}$. Je vous recopie la formule vue en cours car il semble que je l'ai écrite de façon peu lisible au tableau, fatigue de fin de cours en amphi.. :

$$f - u_{\frac{\alpha}{2}} \sqrt{\frac{f(1-f)}{n}} < p < f + u_{\frac{\alpha}{2}} \sqrt{\frac{f(1-f)}{n}}$$

On peut retrouver cette approximation de différentes façons avec les théorèmes limites (théorème central limite sur une suite de Bernoulli, Slutsky, etc.). A noter que $u_{\frac{\alpha}{2}}$ est bien (comme dans le cas 1) toujours le fractile d'ordre $1 - \frac{\alpha}{2}$ de la loi $N(0, 1)$. Pour $\alpha = 0.5$, $u_{\frac{\alpha}{2}} = 1.96$. On trouve, au niveau de confiance de 95%, que $p \in [0.049, 0.1]$.

Exercice 3 : estimation d'une proportion

Cet exercice vise à vous faire prendre conscience de l'impact de la taille de l'échantillon sur l'intervalle de confiance d'une proportion. D'un point de vue statistique, c'est la même formule que précédemment à savoir

$$f - u_{\frac{\alpha}{2}}\sqrt{\frac{f(1-f)}{n}} < p < f + u_{\frac{\alpha}{2}}\sqrt{\frac{f(1-f)}{n}}$$

et on trouve $p \in [0.488, 0.612]$ pour $n = 250$ et $p \in [0.51, 0.58]$ pour $n = 1000$. Ainsi, plus l'échantillon est grand, plus l'intervalle de confiance à 95% se resserre. On donne une fourchette plus précise car plus de personnes ont été interrogées.

Exercice 4 : taille d'échantillon

De même, on va raisonner sur la taille de l'échantillon. Nous voulons trouver un salaire annuel moyen, soit la moyenne m d'une population composée des salaires annuels des ingénieurs en sortie d'école. En faisant l'hypothèse d'une loi normale, on se retrouve dans le cas où l'on cherche à estimer m avec σ^2 connue. Ici l'écart-type est donné, 3500 euros (sur un salaire annuel). Tout l'enjeu de l'exercice consistait à comprendre ce que pouvait signifier le terme marge d'erreur. Sachant que pour le cas 1 vu en cours, l'intervalle de confiance est donné par

$$\bar{x} - u_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} < m < \bar{x} - u_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$$

où u est toujours le fractile d'ordre $1 - \frac{\alpha}{2}$ de la loi $N(0, 1)$. On voit que m est égal à + ou - $u_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$. C'est la marge d'erreur. Il suffit juste de résoudre, pour une marge ϵ donnée,

$$\epsilon = 1.96\frac{\sigma}{\sqrt{n}}$$

et trouver n, le nombre correspondant d'invididus à interroger. Par exemple pour une marge de 500 euros on a

$$n = \frac{1.96^2 * 3500^2}{500^2} = 189.$$

Avec une marge de 100 euros, on trouve que n atteint 4706 individus. Ainsi, pour estimer le salaire annuel, on peut penser qu'une marge de 500 euros est un bon compromis car on aura que 189 individus à interroger. Augmenter la précision fera considérablement augmenter le nombre d'individus à interroger, et donc le budget de l'enquête. Même si ces estimations sont grossières, car elles reposent sur une hypothèse surement irréaliste pour l'écart type, elles permettent de dimensionner une enquête.

Exercice 5 : Indice de Masse Corporelle

L'indice de masse corporelle (IMC) ou Body Mass Index (BMI) est défini par

$$BMI = \frac{poids}{(taille/100)^2}$$

où le poids est mesuré en kg et la taille en cm. On s'intéresse à une population masculine où le BMI a une distribution normale de moyenne $m = 25.33$ et une variance $\sigma^2 = 14.61$. L'obésité est définie par un BMI strictement supérieur à 30.

1) Quel est le pourcentage d'obèses dans cette population ?

On cherche

$$P(BMI > 30) = 1 - P(BMI < 30)$$

On centre et on réduit soit

$$1 - P(U < \frac{30 - 25.33}{\sqrt{14.61}})$$

et U est ainsi la variable qui suit une loi normale $N(0, 1)$. La fonction de répartitition donne

$$P(U < 1.22) = 0.888$$

Donc la proportion recherchée vaut 1 - 0.888 = 0.112. Il y a 11.2% d'obèses dans la population.

2) Dans quel intervalle (centré) se situe le BMI de 95% de la population ?

Ici, c'est une question inhabituelle car l'on connait m et σ. Il s'agit d'un intervalle de fluctuation. On veut un intervalle, que l'on prendra centré sur m, où se situe 95% du BMI soit

$$m \pm u_{\frac{\alpha}{2}}\sigma$$

On trouve $m \in [17.84, 32.82]$.

3) On considère un échantillon de 41 sujets représentatifs de la population masculine. Dans quel intervalle doivent se trouver les valeurs de la moyenne et de la variance du BMI et du pourcentage d'obèses observés sur cette échantillon ?

On va se trouver dans les deux cas avec m et σ^2 connus. Le problème est donc inversé. Ici, nous sommes dans une discipline particulière, l'épidémiologie, et les chercheurs souhaitent vérifier que l'échantillon qu'ils ont choisi est bien représentatif de la population. On cherche en quelque sorte à dimensionner l'étude. Il s'agit donc, dans les formules connues du cours (cas 1 à 4), d'intervertir m et $\bar{x}$. Pour la moyenne observée sur l'échantillon, c'est-à-dire $\bar{x}$, nous obtenons

$$\bar{x} \in m \pm u_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

et on obtient $\bar{x} \in [24.16, 26.50]$. En ce qui concerne la variance de l'échantillon, on peut utiliser la statistique

$$T = \frac{1}{n} \sum_{i=1}^{n} \big(X_i - m\big)^2$$

et l'on sait que $\frac{nT}{\sigma^2}$ suit une loi χ^2_n.

le fractile k_2 vaut 60.56 et le fractile k_1 vaut 25.22 (pour 41 degrés de liberté, voir la valeur sur http ://rfv.insa-lyon.fr/ jolion/STAT/node142.html). J'obtiens en résultat $\sigma^2 \in [9.89, 23.75]$.

4) On tire au sort 10 sujets de la population. Les valeurs du BMI observées sont les suivantes :

$$(20, 21, 22, 25, 27, 28, 29, 31, 32, 33).$$

Donnez l'estimation et l'intervalle de confiance à 95% de :

- La moyenne du BMI et sa variance,
- La proportion d'obèses.

On revient à une question très classique, comme vue dans le cours et les exercices précédents. On peut appliquer les formules du cours. Pour la moyenne on a

$$\bar{x} - t\frac{s}{\sqrt{n-1}} < m < \bar{x} - t\frac{s}{\sqrt{n-1}}$$

et l'on trouve $m \in [23.47, 30.14]$. Pour la variance on a

$$\frac{ns^2}{l_2} < \sigma^2 < \frac{ns^2}{l_1}$$

et la valeur trouvée est $\sigma^2 \in [10.2, 72.44]$. Pour la proportion, on applique la formule désormais bien connue et on trouve $p \in [0.016, 0.58]$.