Paul-Antoine Chevalier, Frédéric Pascal et Hugo Harari-Kermadec 7 octobre 2010

Statistiques

TD 2 : Estimation et intervalles de confiance

L3, ENS-Cachan, 2010-2011

2.1 Le prix des logements à Brest

On a collecté des données sur les prix des logements à Brest [1]. On a représenté les données brutes des prix pour 249 observations.

```
  [1] 324 378 251 433 338 127 197 222 211 496 269 183 302 227 417 291 314 155
 [19] 108 108 602 360 290 300 276 271 234 255  68 298 130 216 291 216 108 141
 [37] 140 140 141 118 162 195 249  91 162  99 205 247 211 180 211 300 223  64
 [55]  64  76 222 248 281 218 178 432  77  81 173 259 217 162 437 181 213 281
 [73] 165 183 146 317 290 216  97  74 110  74 191 258 216 167 167 196 203  91
 [91] 255 155  76 331 159 159 271 271 290 226 149 149 388 249 245 226 173  57
[109] 192 162 130  92  92  86 114 108 110 205 108 300 291 162 108 377 124 159
[127] 323 344 237 217 582 252 270 216 173 140 131 119 178 274 140 171 287 226
[145] 405 274 301 291 259 248 183 147 210 198 230 220 170 108 284 151 195 150
[163] 348  87 170 200 151  91  65 292 108 162 105 266 164  75 162 156 158 140
[181] 228 256 313 151 354 247 289 140 541 251 114 130 201 216 149 173 167 103
[199] 123 210  95 216 356 210 183 306 205 254 151 183 209 178 146 253  81  81
[217] 263 367 226 141 108 243 208 200 151  54 128 106  95  95 126  90 232 173
[235]  95 143 143 324 152 107  71  69 165 132 200 130  85 137 136
```

On tente de résumer l'information contenue dans les données par quelques statistiques descriptives. On remarque que la moyenne empirique vaut $\hat{m} = \frac{1}{N}\sum x_i = 200.87$ et l'écart-type empirique vaut $\hat{\sigma} = \sqrt{\frac{1}{N-1}\sum(x_i - \hat{m})^2} = 94.37$

☞ **Q1**

Dans un premier temps, on fait l'hypothèse que le prix des logements suit une loi normale. D'après les informations ci-dessus, quelles seraient alors les valeurs attendues du premier quartile, de la médiane et du troisème quartile ?

☞ **Q2**

On observe dans les données que le premier quartile vaut 136.0, la médiane 183.0 et le troisième quartile 254.0. Qu'en pensez-vous ?

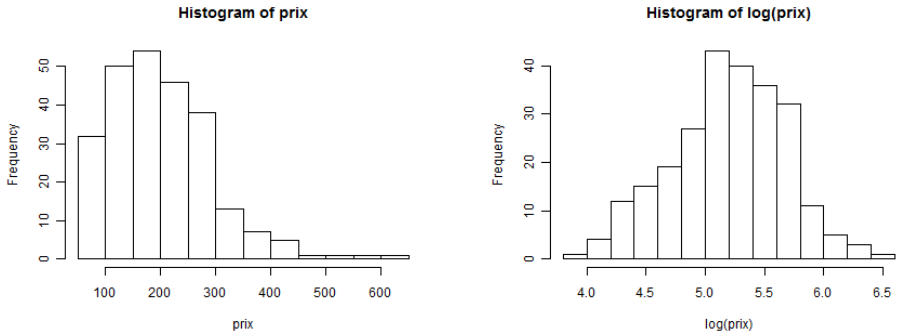
1. Les données proviennent du site d'Emmanuel Flachaire http://www.vcharite.univ-mrs.fr/PP/flachaire/teaching/cours_m1_econometrie_appliquee.html. Elles ont été exploitées dans l'article : "Prix des logements et autocorrélation spatiale : une approche semi-paramétrique" (E. Flachaire, I. Ahamada et M. Lubat), Economie Publique, 2007.

☞ **Q3**

On propose de faire l'hypothèse que les prix suivent une loi log-normale. La moyenne empirique du log des prix vaut 5.195 et l'écart type estimé 0.4728. Sous cette hypothèse, quelles sont les valeurs attendues du premier quartile, de la médiane et du troisième quartile ?

☞ **Q4**

On observe que la médiane vaut 5.209, le premier quartile 4.913 et le troisième quartile vaut 5.537. Par ailleurs, on représente les histogrammes (estimation de la densité) du prix et du log des prix.



Qu'en concluez-vous ?

2.2 La correction d'un paquet de copie

Le chargé de TD corrige un paquet de 30 copies. Les copies sont notées entre 0 et 20 :

```
 [1] 13  9 12 10  7 14 15  7 11 13 11 13 14 12  8
[16] 14 17 13 16 17 16 14 18 13 13 17  8 10 11 20
```

On note X_n la note attribuée à la énième copie. On suppose que les notes sont indépendantes et identiquement distribuées (iid). On note m l'espérance de X_n et σ^2 sa variance. Après chaque copie, le chargé de TD calcule la moyenne empirique $\overline{X}_n$.

☞ **Q1**

Quel est la variance de $\overline{X}_n$?

☞ **Q2**

Peut-on estimer la variance σ^2 de X_i lorsque l'on a qu'une seule observation ? Pourquoi ?

☞ **Q3**

On propose d'estimer la variance par $S^2 = \frac{1}{n}\sum(X_i - \overline{X}_n)^2$. Montrez que cet estimateur est biaisé. On pourra montrer dans un premier temps que

$$S^2 = \frac{1}{n}\sum(X_i - m)^2 - (\overline{X} - m)^2$$

☞ **Q4**

Proposez un estimateur sans biais de la variance.

☞ **Q5**

Représenter sur un graphique l'écart-type de $\overline{X}_n$ en fonction de n. Qu'en déduisez vous ?

2.3 Guitare électrique

Désireux d'acheter une guitare électrique, vous avez collecté des prix dans une boutique : 390, 460, 650, 410, 270 et 780 euros. Vous cherchez θ_0, l'espérance du prix d'une guitare. Un de vos camarades, plus malin que vous, a collecté 300 annonces sur internet. Il obtient un prix moyen de 550 euros et un écart-type empirique de 300 euros.

☞ **Q1**

A partir de deux échantillons supposés constitués d'observations indépendantes et identiquement distribuées proposez deux estimations θ_1 et θ_2 de l'espérance du prix d'une guitare.

Quelles propriétés possèdent ces estimateurs ? Précisez les hypothèses nécessaires à chaque propriété.

☞ **Q2**

Votre professeur de Statistique vous dit qu'il y a un moyen d'améliorer encore la précision de l'estimation de θ_0 en combinant votre estimation et celle de votre camarade. Il vous propose de calculer la moyenne arithmétique des deux estimations en les pondérant donc chacune par 0.5 : $\hat{\theta}_* = 0.5\hat{\theta}_1 + 0.5\hat{\theta}_2$, où $\hat{\theta}_1$ et $\hat{\theta}_2$ représentent respectivement votre estimation de θ_0 et celle de votre camarade. Calculer la variance de cette nouvelle estimation de l'espérance du prix.

☞ **Q3**

Votre chargé de TD vous fait remarquer que votre professeur de statistique pourrait être plus malin et qu'en choisissant mieux la pondération de ces deux estimations, on peut encore améliorer la précision du résultat. Soient a et $1-a$ un jeu de pondérations : $\hat{\theta}_{**} = a\hat{\theta}_1 + (1-a)\hat{\theta}_2$.

Donnez, en fonction de a, la variance de ce nouvel estimateur de l'espérance du prix. Quelle est la valeur de a qui minimise cette variance ?

☞ **Q4**

Construisez un intervalle de confiance asymptotique à 95% pour l'esérance du prix à partir des deux échantillons initiaux (avec $n = 6$ pour l'un et $n = 300$ pour l'autre) ainsi qu'à partir de l'échantillon complet regroupant les deux. Qu'observez-vous ?

2.4 Sondage Bush

Nous avons des données issues d'une enquête demandant aux Américains s'ils approuvent la politique menée par le président Bush. Il y a 1500 répondants contactés par un sondage aléatoire. On obtient que 35% des répondants soutiennent la politique de Bush. Construisez un intervalle de confiance à 95 % autour de la valeur estimée.

2.5 Combien de parisiens portent des lunettes

(d'après P. Ardilly et Y. Tillé, Exercices corrigés de méthodes de sondage)
Quelle taille d'échantillon faut-il retenir pour connaître à deux points de pourcentage près (au plus) et avec 95 chances sur 100, la proportion de Parisiens qui portent des lunettes ? On suppose que chaque Parisien a la même probabilité d'être sondé, et que les individus sont tirés avec remise.