Eric Vandermolen

3ème année en ingénieur civil

Eléments de statistique

Partie n° 2 du projet personnel

7 décembre 2016

Estimation

(a) Nous savons par définition que le biais et la variance de l'estimateur m_x sont donnés par :

$$V(m_x) = E\left[(m_x - E(m_x))^2\right]$$

et

$$B(m_x) = |E(m_x) - \mu|$$

avec m_x : la moyenne de l'échantillon et μ : la moyenne de la population.
En faisant tourner le code, nous trouvons que $V(m_x) = 0.5260$ et que $B(m_x) = 0.0767$.

(b) On sait par définition que le biais et la variance de l'estimateur $median_x$ sont donnés par :

$$V(median_x) = E\left[(median_x - E(median_x))^2\right]$$

$$B(median_x) = |E(median_x) - \mu|$$

avec $median_x$: la médiane de l'échantillon.
En faisant tourner le code, nous trouvons que $V(median_x) = 0.6446$ et que $B(median_x) = 0.1017$.

(c) En augmentant le nombre de nos échantillons, nous obtenons pour la moyenne : $V(m_x) = 0.1698$ et que $B(m_x) = 0.0230$. Concernant la médiane, on a que $V(median_x) = 0.1935$ et que $B(median_x) = 0.0492$.
Maintenant analysons les résultats obtenus : à nombre d'échantillons identiques, nous remarquons que la variance pour chacun des estimateurs sont assez proches l'une de l'autre contrairement au biais qui est moins important pour la moyenne m_x pour chacun des 2 tailles d'échantillons. D'après la théorie nous savons que la moyenne m_x est un meilleur estimateur que la médiane $median_x$. En recommançant plusieurs fois les calculs, nous remarquons que le biais pour la médiane est parfois moins élevée que celle de la moyenne. Les valeurs étant soumis à l'aléatoire, ceci est logique mais au fur et à mesure qu'on augmente la taille de l'échantillon, les valeurs obtenues tendent vers celles décrites par la théorie. Et m_x tendrait vers la moyenne de la population contrairement à $median_x$ qui garderait un certain biais.

(d) L'intervalle de confiance étant à 95%, nous avons donc que $\alpha = 0.05$.

(i) En utilisant la loi de Student :
nous construisons une loi de Student à $n - 1 = 19$ ddls, l'intervalle de confiance est :

$$IC_{Student_{0.95}}(\mu) = \left[m_x - t_{1-\alpha/2}\frac{s_{19}}{\sqrt{20}}, m_x + t_{1-\alpha/2}\frac{s_{19}}{\sqrt{20}}\right]$$

(ii) En utilisant la loi de Gauss :
nous avons comme intervalle :

$$IC_{Gauss_{0.95}}(\mu) = \left[m_x - u_{\alpha/2}\frac{s_{19}}{\sqrt{20}}, m_x + u_{\alpha/2}\frac{s_{19}}{\sqrt{20}}\right]$$

Les valeurs de $t_{1-\alpha/2}$ et de $u_{\alpha/2}$ étant disponible dans des tables, en faisant tourner le code, on tourne autour de 96 pour Student et 94 pour Gauss, les 2 valeurs étant proches de α. La loi de student convergent vers la loi de Gauss pour $n > 30$, les valeurs obtenues sont donc assez proches l'une de l'autre.

Afin de vérifier s'il était raisonable de supposer la variable parente Gausienne, il suffit de vérifier si 68% des résultats finaux des étudiants appartiennent à l'intervalle : $[\mu - \sqrt{\sigma}, \mu + \sqrt{\sigma}]$,

μ et $\sqrt{\sigma}$ étant respectivement la moyenne et l'écart-type des résultats finaux des étudiants. En vérifiant dans matlab combien de résultats appartiennent dans l'intervalle, on trouve que 68.92% des résultats se trouvent dans l'intervalle, il était donc résonable de supposer la variable parente Gausienne.

Tests d'hypothèse

On a les hypothèses suivantes :

— H_0 : un quart des étudiants ont obtenu une note inférieur à 10 et l'hypothèse alternative,

— H_1 : plus d'un quart des étudiants ont obtenu une note inférieur à 10.

Nous supposons que

$$f \sim \mathcal{N}\left(p, \left(\sqrt{\frac{p(1-p)}{n}}\right)^2\right) = 1 - \alpha$$

$$\Leftrightarrow f \sim \mathcal{N}\left(0.25, \left(\sqrt{\frac{0.25 \times 0.75}{20}}\right)^2\right) = 0.95$$

p étant la proportion d'étudiant ayant obtenu une note inférieur à 10 et f l'estimateur de la côte des étudiants

Pour déterminer un intevalle I_{H_0} qui ne contient que les valeurs de f favorable à H_0, il faut déterminer ϵ tel que

$$P\left(f \leq p + \epsilon\right) = 1 - \alpha$$

$$\Leftrightarrow P\left(f \leq 0.25 + \epsilon\right) = 0.95$$

En exprimant la loi de Gauss sous la forme centrée réduite, nous obtenons

$$P\left(\mathcal{Z} \leq \frac{\epsilon}{\sigma}\right) = 0.95$$

En regardant dans les tables, on voit que 0.95 correspond approximativement à la valeur 1.65. Et donc que $\epsilon = 0.1598$.

Et donc que

$$I_{H_0} =]-\infty, 0.4098]$$

(a) En faisant le test d'hypothèse, nous voyons que dans 4 cas sur 100, l'Ulg rejete l'hypothèse H_1. Et en tournant le code plusieurs fois, nous voyons que celle-ci a tendance à être proche au seuil de signification α mais très souvent légrement inférieur à celui-ci.

(b) En faisant le test d'hypothèse, nous voyons que dans 21 cas sur 100, un article de la gazette sera publié. La valeur obtenue est plus importante que celle obtenue au point précédent. Celle-ci est logique car un article est publié si au moins un institut de sondage externe critique le cours de probabilité, étant donné qu'ils sont 6, il y a donc 6 fois plus de chance en théorie que l'hypothèse soit rejeté par les instituts de sondage. Pour ce test d'hypothèse-ci, nous obtenons que les instituts de sondage externe ont 5.25 fois plus de chance de publié un article que l'Ulg.

(c) Plusieurs moyens peuvent être mis en oeuvre afin de réduire l'avantage des différentes institus de sondage externe :

- Nous pouvons par exemple diminuer le seuil de signifinication α pour les instituts de sondage. Dans ce cas, il y aura donc moins de chance que l'hypothèse H_0 soit rejeté et après d'ajuster la valeur de α tel que les probabilités soient égales.
- Nous pouvons également augmenter la taille des échantillons pour les instituts de sondage afin qu'ils aient moins de chance de rejetter l'hypothèse H_0 et après d'ajuster la taille de l'échatillon tel que les probabilités soient égales.
- En utilisant des conditions plus strictes concernant la publication d'un article, c'est-à-dire qu'un article est publié si et seulement un certain nombre des 6 instituts de sondage externe ont rejeté sumultanément l'hypothèse H_0.
- En fusionnant les 6 instituts de sondage en une seule entitée.

Code matlab

```
function statProjet2

%% Question 1: Estimation

% generation de 100 echantillons iid de taille 20:

map = xlsread('Proba1ereSess20122013.xls');
map = int8(map);

meanStudent = mean(map,2);
meanMeanStudent = mean(meanStudent)

varIid20 = zeros(100,1);
iid20MeanStudent = zeros(100,20);
for a=1:100,
    iid20MeanStudent(a,:) = datasample(meanStudent,20);
    varIid20(a,1) = var(iid20MeanStudent(a,:));
end

% (a) moyenne: biais et variance:

meanIid20 = zeros(100,1);
for a=1:100,
    meanIid20(a,1) = mean(iid20MeanStudent(a,:));
end

meanMeanIid20 = mean(meanIid20);
varMeanIid20 = var(meanIid20)
biaisMeanIid20 = meanMeanIid20 - meanMeanStudent

% (b) mediane: biais et variance:

medianIid20 = zeros(100,1);
for a=1:100,
```

```
    medianIid20(a,1) = median(iid20MeanStudent(a,:));
end

meanMedianIid20 = mean(medianIid20);
varMedianIid20 = var(medianIid20)
biaisMedianIid20 = meanMediantIid20 - meanMeanStudent

% (c) meme chose mais avec echantillons iid de taille 50:

iid50MeanStudent = zeros(100,50);
for a=1:100,
    iid50MeanStudent(a,:) = datasample(meanStudent,50);
end

meanIid50 = zeros(100,1);
for a=1:100,
    meanIid50(a,1) = mean(iid50MeanStudent(a,:));
end

meanMeanIid50 = mean(meanIid50);
varMeanIid50 = var(meanIid50)
biaisMeanIid50 = meanMeanIid50 - meanMeanStudent

medianIid50 = zeros(100,1);
for a=1:100,
    medianIid50(a,1) = median(iid50MeanStudent(a,:));
end

meanMedianIid50 = mean(medianIid50);
varMedianIid50 = var(medianIid50)
biaisMedianIid50 = meanMedianIid50 - meanMeanStudent

% (d) intervalle de confiance a 95%

s_19 = zeros(100,1);
for a=1:100,
    s_19(a,1) = sum((meanIid20(a,1)-meanMeanStudent).^2*sqrt(20/19));
end
s_19
% i. loi de student a 19 ddl:

borneInfS = zeros(100,1);
borneSupS = zeros(100,1);
intervalleStudentOk = 0;
for a=1:100,
    borneInfS(a,1) = meanIid20(a,1) - 2.093.*s_19(a,1)/sqrt(20);
```

```
    borneSupS(a,1) = meanIid20(a,1) + 2.093.*s_19(a,1)/sqrt(20);
    if (borneInfS(a,1) < meanMeanStudent && meanMeanStudent < borneSupS(a,1)),
        intervalleStudentOk = intervalleStudentOk + 1;
    end
end

intervalleStudentOk


% ii. loi de Gauss:

borneInfG = zeros(100,1);
borneSupG = zeros(100,1);
intervalleGaussOk = 0;
for a=1:100,
    borneInfG(a,1) = meanIid20(a,1) - 1.96.*s_19(a,1)/sqrt(20);
    borneSupG(a,1) = meanIid20(a,1) + 1.96.*s_19(a,1)/sqrt(20);
    if (borneInfG(a,1) < meanMeanStudent && meanMeanStudent < borneSupG(a,1)),
        intervalleGaussOk = intervalleGaussOk + 1;
    end
end

intervalleGaussOk

% hypothese parente Gauss correct?:


normalOK = 0;
[a ~] = size(meanStudent);
for i=1:a,
    if mean(meanStudent)-std(meanStudent) < meanStudent(i) && mean(meanStudent)+s
        normalOK = normalOK + 1;
    end
end

normalOK = normalOK/a

%% Question 2: Tests d'hypothese:

% generation de 7 echantillons iid de taille 20 (100x):

echInstituts = zeros(100,7,20);
for a=1:100,
    for b =1:7,
        echInstituts(a,b,:) = datasample(meanStudent,20);
    end
```

```
end

epsillon = 1.65*sqrt(0.25*0.75/20);
borneSup = epsillon + 0.25

% (a) rejet de l'hypothese par l'Ulg:

meanUlg(100,1) = 0;
rejetUlg = 0;
for a = 1:100,
    for b = 1:20,
        if (echInstituts(a,1,b) < 10),
            meanUlg(a,1) = meanUlg(a,1) + 1;
        end
    end
    meanUlg(a,1) = meanUlg(a,1)./20;
    if (meanUlg(a,1) > borneSup)
        rejetUlg = rejetUlg + 1;
    end
end

rejetUlg

% (b) rejet de l'hypothese par un institut de sondage externe:

meanInstitut(100,6) = 0;
publication = zeros(100,1);
for a=1:100,
    for b=1:6,
        for c=1:20,
            if (echInstituts(a,b+1,c) < 10),
                meanInstitut(a,b) = meanInstitut(a,b) + 1;
            end
        end
        meanInstitut(a,b) = meanInstitut(a,b)./20;
        if (meanInstitut(a,b) > borneSup),
            publication(a,1) = 1;
        end
    end
end

publication = mean(publication)*100
probaAvantage = publication/rejetUlg
```