Projet Statistique 2

Mehdi Sauvage

December 8, 2016

1 Estimation

(a)

Pour chaque échantillons idd, le calcul de la variance et su biais pour la moyenne des échantillons a été exécuté comme définit ci-dessous.

Définition de la variance :

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

Définition du biais:

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

Dans le cas des 100 échantillons de 20 étudiants, la variance vaut 0.340 et le biais est égal à 0.16.

(b)

Définition de la variance :

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

Définition du biais:

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

Dans le cas des 100 échantillons de 20 étudiants, la variance vaut 0.499 et le biais est égal à 0.276.

(c)

Considèreront le cas ou les échantillons sont de tailles 50. Il en suit une variation de la variance et du biais. en effet,

X	échantillons 20	échantillons 50
$V(m_x)$	0.16	0.1141
$B(m_x)$	0.1666	0.0049
$V(med_x)$	0.4899	0.2177
$B(med_x)$	0.3809	0.2742

Nous pouvons constater que l'augmentation de la taille des échantillons fait diminuer toutes les valeurs de biais et de variance. Intuitivement, cela a du sens d'imaginer que en augmentant l'information contenue dans chaque échantillons, nous nous rapprochons des valeurs réelles de la population.

Notons, que le biais des moyennes tendra vers la moyenne de la population alors que le biais de la médiane ne tendra pas indéfiniment vers la médiane de la population.

(d)

d.i - Loi de Student

L'intervalle de confiance se calcule comme suit :

$$m_x - t_{1-\alpha/2}\frac{s_{n-1}}{\sqrt{n}} < x < m_x + t_{1-\alpha/2}\frac{s_{n-1}}{\sqrt{n}}$$

Notons que la valeur de $t_{1-\alpha/2}$ se trouve dans les tables.

d.ii - Loi de Gauss

L'intervalle de confiance se calcule comme suit :

$$m_x - u_{\alpha/2}\frac{\sigma}{\sqrt{n}} < x < m_x + u_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Notons que la valeur de $\mu_{\alpha/2}$ se trouve dans les tables.

d. - Explications

Le nombre de dégrées de liberté dans les deux cas est $n - 1$ soit 19. De plus, nous remarquons que les résultat donnée par la loi Student et la loi de Gauss sont fort similaire.

En effet, la théorie dit que pour des échantillons suffisamment grand, les intervalles se confond de plus en plus ($n > 30$). Le cas ou $n = 20$ n'étant pas très éloigné, la similitude entre les résultat semble cohérent.

La moyenne de la population a été comprise 96 fois sur 100 dans les intervalles généré. Nous remarquons que cette valeur est proche de $1 - \alpha$.

2 Test d'hypothèse

Soit les deux hypothèse H_0 et H_1 définit comme suit:

- 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 ,un estimateur des résultats des étudiants, est une variable qui suit une lois normal.

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

ou p étant la proportion d'étudiant avec une moyenne ¡ 10.
il faut déterminer ϵ tel que l'intervalle pour H_0 ait des valeurs de f favorable.

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

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

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

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

Pour trouver $\epsilon = 0.1598$ nous devons chercher dans la table de gauss.

(a)

L'ULg rejet l'hypothèse que au moins un quart des étudiants ait obtenu une note final ¡ 10/20 dans 4 cas. Notez que le nombre de cas de rejet s'approche de la valeur de α.

(b)

Les instituts de sondage rejet l'hypothèse que au moins un quart des étudiant ai obtenu une note final ¡ 10/20 dans 19 cas. Notez que le nombre de cas ou l'hypothèse est rejeter est plus importante de dans pour les autorités de l'ULg.Ceci est du au faite que les instituts de sondage sont plus nombreux et on donc plus de chances de se trouver face a un échantillons à rejeter.

(c)

Pour minimiser l'influence du problème énoncé au point (c) nous pouvons donner le même échantillons a tout les instituts de sondage. Nous pourrions aussi alléger les conditions de rejet pour les instituts de sondage en diminuant α. Notons aussi qu'en augmentant le nombre d'étudiant par échantillons, les instituts de sondage (tout comme les autorité de l'ULg) serait aussi moins souvent confronté au rejet de l'hypothèse H_0

Code

```
function [biais_mx , var_mx] = Q1a( ech , mean_pop ,
    var_pop )

    %Definition de la taille de l'echantillon
    size_n = size(ech(1,:));
    n = size_n(2);

    mean_ech_n = zeros(100,1);

    %Calcul des moyennes des echantillons
    for i = 1 : 100
        mean_ech_n(i) = mean(ech(i,:));
    end
    %Moyenne des moyennes
    mean_mean_ech_n = mean(mean_ech_n(:));

    %Calcul de la variance
    var_mx = var(mean_ech_n);

    %Calcul du biais
    biais_mx = mean_mean_ech_n - mean_pop ;

end
```

```
function [biais_medx , var_medx] = Q1b( ech , mean_pop ,
    var_pop )

    %Definition de la taille de l'echantillon
    size_n = size(ech(1,:));
    n = size_n(2);

    med_ech_n = zeros(100 , 1);

    %Calcul des medianes des echantillons
    for i = 1 : 100
        med_ech_n(i) = median(ech(i,:));
    end


    %Calcul de la moyenne des  m dianes
    mean_med_ech_n = mean(med_ech_n(:));

    %Calcul de la variance
    var_medx = var(med_ech_n);

    %Calcul du biais
    biais_medx = mean_med_ech_n - mean_pop ;

end
```

```
function [borne_inf_norm ,borne_sup_norm ,borne_inf_stud ,borne_sup_stud ,stud ,norm] ...
    = Q1d(ech , mean_pop)

    %Definition de la taille de l'echantillon et initialisations
    size_tmp = size(ech(1,:));
    n = size_tmp(2);

    borne_inf_norm = zeros(1,n);
    borne_sup_norm = zeros(1,n);
    borne_inf_stud = zeros(1,n);
    borne_sup_stud = zeros(1,n);

    mean_ech_n = zeros(n,1);
    ET_ech_n = zeros(n,1);

    stud = 0;
    norm = 0;

    for i = 1 : 100

        mean_ech_n(i) = mean(ech(i,:));
        ET_ech_n(i) = std(ech(i,:));

        %Lois Normal - calcul des bornes de l'ic
        borne_inf_norm(i) = mean_ech_n(i) - 1.96 * (ET_ech_n(i) / sqrt(n));
        borne_sup_norm(i) = mean_ech_n(i) + 1.96 * (ET_ech_n(i) / sqrt(n));

        %Loi student - calcul des bornes de l'ic
        k = 2.093;
        s_n_1 = sqrt(n / (n -1)) * ET_ech_n(i) ;

        borne_inf_stud(i) = mean_ech_n(i) - k * (s_n_1 / sqrt(n));
        borne_sup_stud(i) =mean_ech_n(i) + k * (s_n_1 / sqrt(n));


        %Verification des ic avec la moyenne de la population
        if ( mean_pop > borne_inf_stud(i) && mean_pop < borne_sup_stud(i))
            stud = stud + 1;
        end

        if ( mean_pop > borne_inf_norm(i) && mean_pop < borne_sup_norm(i))
```

```
                norm = norm + 1;
            end

        end

    end
```

```
function [ulg_rejet , ulg_mean] = Q2a(ech)

%initialisation de la borne suprieur
eps = 0.1598;
sup = eps + 0.25;

ulg_mean = zeros(100,1);
ulg_rejet = 0;

for i = 1 : 100
    for j = 1 : 20
        %verif condition d'hypothese ( < 10 )
        if(ech(i,1,j) < 10)
            ulg_mean(i) = ulg_mean(i) + 1;
        end
    end

    %calcul de la proportion de cas <10
    ulg_mean(i) = ulg_mean(i)/20;

    %calcul de la proportion de cas rejet
    if (ulg_mean(i) > sup)
        ulg_rejet = ulg_rejet + 1;
    end
end

end
```

```
function [inst_rejet_ , inst_mean] = Q2b(ech)

%initialisation de la borne sup rieur
eps = 0.1598;
sup = eps + 0.25;

inst_mean = zeros(100,6);
inst_rejet = zeros(100,1);

for i = 1 : 100
    for j = 2 : 6
        for k = 1 : 20
            %verif condition d'hypothese ( < 10 )
            if(ech(i,j,k) < 10)
                inst_mean(i , j) = inst_mean(i,j) + 1;
            end
        end

    inst_mean(i,j) = inst_mean(i,j)/20;

    if (inst_mean(i,j) > sup)
        inst_rejet(i) = 1;
    end
    inst_rejet_ = mean(inst_rejet) * 100;
end

end
```