**Université du Sud Toulon-Var**

Master Enseignement **1**ère année

TD 7 - Statistique : Estimations et tests d'hypothèses.

Quelques notions sur les estimations et les tests d'hypothèses

I. ESTIMATIONS

On considère une expérience à laquelle est attachée une variable aléatoire $X$ dont la loi dépend d'un paramètre inconnu. On répète alors cette expérience un certain nombre $n$ de fois, de façon indépendante, et si $(\omega_1, \ldots, \omega_n)$ est la suite des $n$ résultats obtenus, on définit $n$ variables aléatoires $X_1$, $X_2$, ..., $X_n$ indépendantes et toutes de même loi que $X$ par

$$X_i(\omega) = X(\omega_i).$$

On sera alors amené à déduire de l'observation des ces résultats deux types d'estimations :

**1**. Des **estimations ponctuelles**, dans lesquelles on cherche à déterminer une nouvelle variable aléatoire $Y_n^* = f(X_1, X_2, ...X_n)$ appelée **estimateur**, d'espérance le paramètre inconnu, et de variance qui tend vers zéro quand $n$ tend vers l'infini. On prendra alors pour estimation ponctuelle du paramètre la valeur réelle prise par $Y_n^*$ lors de l'observation : $f(x_1, x_2, ..., x_n)$, avec $x_i = X(\omega_i) = X_i(\omega)$ étant la $i$ème valeur observée de l'échantillon.

**2**. Des **estimations par intervalle de confiance**, pour lesquelles on se donne un réel $\alpha > 0$ "petit", appelé **seuil de confiance** ou **risque**, et on cherche à déterminer deux statistiques $Y_n^* = f(X_1, X_2, ..., X_n)$ et $Z_n^* = g(X_1, X_2, ...X_n)$ telles que si on note $\theta$ le paramètre inconnu, on aura

$$\mathbb{P}(\theta < Y_n^*) = \mathbb{P}(\theta > Z_n^*) = \alpha/2\,,$$

i.e., on cherche à rejeter les valeurs de $\theta$ trop petites ou trop grandes.

On aura alors

$$\mathbb{P}(Y_n^* \leq \theta \leq Z_n^*) = \mathbb{P}(\theta \leq Z_n^*) - \mathbb{P}(\theta < Y_n^*) = (1 - \alpha/2) - \alpha/2 = 1 - \alpha\,.$$

Donc, si

$$I = [Y_n^*,\, Z_n^*]$$

on aura $\mathbb{P}(\theta \in I) = 1 - \alpha$.

On dira alors que $I = [Y_n^*,\, Z_n^*]$ est un **intervalle de confiance** pour $\theta$ avec la confiance $1 - \alpha$.

Pour l'application numérique, on remplacera $Y_n^*$ et $Z_n^*$ respectivement par $f(x_1, ..., x_n)$ et $g(x_1, ..., x_n)$.

**1. Exemples usuels d'estimations ponctuelles**

**1.a. Estimation de l'espérance d'une variable aléatoire**

Soit $m$ l'espérance inconnue d'une variable aléatoire $X$.

On considère alors la statistique

$$M_n^* = \frac{X_1 + \ldots + X_n}{n}\,. \tag{1}$$

On a alors $\mathbb{E}(M_n^*) = m$ et $\mathrm{Var}(M_n^*) = \frac{\mathrm{Var}(X)}{n} \to 0$ quand $n \to +\infty$.

$M_n^*$ est donc un estimateur de $m$.

Si on dispose d'une série de mesures (échantillon) $x_1$, $x_2$, ... $x_n$ pour $X$, une estimation ponctuelle (numérique) de $m$ sera alors

$$m_n^* := \frac{x_1 + x_2 + \ldots + x_n}{n}$$

parfois aussi notée $\bar{x}$.

**1.b. Estimation d'une variance d'une variable aléatoire dont l'espérance $m$ est connue.**

On considère la statistique

$$V_n^* = \frac{1}{n}\left((X_1 - m)^2 + \cdots + (X_n - m)^2\right) . \tag{2}$$

On a alors $\mathbb{E}(V_n^*) = \mathrm{Var}(X)$ et on montre que $\mathrm{Var}(V_n^*) \to 0$ quand $n \to \infty$.

Un estimateur de $\mathrm{Var}(X)$ est donc $V_n^*$.

Si on dispose d'une série de mesures (échantillon) $x_1$, $x_2$, ... $x_n$ pour $X$, une estimation ponctuelle (numérique) de $\mathrm{Var}(X)$ sera alors

$$v^* = \frac{1}{n}\left((x_1 - m)^2 + (x_2 - m)^2 + \ldots + (x_n - m)^2\right) .$$

On déduit aussi aisément de ceci une estimation ponctuelle de l'écart type.

**1.c. Estimation de la variance d'une variable aléatoire dont l'espérance $m$ est inconnue**

Soit la variable aléatoire

$$V_n^{*\prime} = \frac{1}{n-1}\left((X_1 - M_n^*)^2 + \ldots + (X_n - M_n^*)^2\right) .$$

On montre que $\mathbb{E}(V_n^{*\prime}) = \mathrm{Var}(X)$ et $\mathrm{Var}(V_n^{*\prime}) \to 0$ quand $n \to +\infty$.

Un estimateur de $\mathrm{Var}(X)$ est donc $V_n^{*\prime}$.

Si on dispose d'une série de mesures (échantillon) $x_1$, $x_2$, ... $x_n$ pour $X$, une estimation ponctuelle (numérique) de $\mathrm{Var}(X)$ sera alors

$$v_n^{*\prime} = \frac{1}{n-1}\left((x_1 - m^*)^2 + \ldots + (x_n - m^*)^2\right) .$$

**1.d. Estimation d'une proportion**

Dans ce cas là, la variable aléatoire $X$ est la répétition d'une loi de Bernoulli $\mathcal{B}(1, p)$. L'estimateur est

$$M_n^* = \frac{X_1 + \ldots + X_n}{n} .$$

On a alors $\mathbb{E}(M_n^*) = p$ et $\mathrm{Var}(M_n^*) = \frac{1}{n^2}(np(1-p)) \to 0$ $n \to +\infty$.

$M_n^*$ est donc un estimateur de $p$.

Si on dispose d'une série de mesures (échantillon) $x_1$, $x_2$, ... $x_n$ pour $X$, une estimation ponctuelle (numérique) de $m$ sera alors

$$p^* := \frac{x_1 + x_2 + \ldots + x_n}{n} .$$

**2. Estimation par intervalle de confiance des paramètres d'une loi normale.**

On considère ici que la loi de la variable aléatoire $X$ est une loi normale $\mathcal{N}(m, \sigma)$. Le paramètre à estimer sera soit $m$ soit $\sigma$.

**2.a. Estimation par intervalle de confiance de $m$ si $\sigma$ est connu**

Puisque $X_i \sim \mathcal{N}(m, \sigma)$, alors $M_n^* \sim \mathcal{N}(m, \frac{\sigma}{\sqrt{n}})$ (le montrer)

Soit la variable aléatoire

$$R_n^* = \frac{M_n^* - m}{\sigma/\sqrt{n}} .$$

Alors $R_n^* \sim \mathcal{N}(0, 1)$.

Pour $h \in \mathbb{R}_+^*$, on a alors

$$\{|R_n^*| \leq h\} = \{-h \leq R_n^* \leq h\} = \{-h \leq \frac{M_n^* - m}{\sigma/\sqrt{n}} \leq h\} = \{M_n^* - h\frac{\sigma}{\sqrt{n}} \leq m \leq M_n^* - h\frac{\sigma}{\sqrt{n}}\}$$

Si on fixe $h$ tel que $\mathbb{P}(|R_n^*| \leq h) = 1 - \alpha$, on aura alors trouvé un intervalle de confiance pour $m$ avec la confiance $\alpha$ en prenant

$$Y_n^* = M_n^* - h\frac{\sigma}{\sqrt{n}} \quad \text{et} \quad Z_n^* = M_n^* + h\frac{\sigma}{\sqrt{n}} .$$

Si on dispose d'une série de mesures (échantillon) $x_1$, $x_2$, ... $x_n$ pour $X$, une estimation (numérique) par intervalle de confiance de $m$ sera alors

$$I = [m^* - h\frac{\sigma}{\sqrt{n}},\, m^* + h\frac{\sigma}{\sqrt{n}}]$$

où

$$m^* := \frac{x_1 + x_2 + \ldots + x_n}{n}\,.$$

On résume le résultat de la façon suivante :

*Estimation par intervalle de confiance de $m$ si $\sigma$ est connu*

– Seuil de confiance : $\alpha$ (ou confiance $1-\alpha$)

– Intervalle de confiance (aléatoire) :

$$I = [M_n^* - h\frac{\sigma}{\sqrt{n}},\, M_n^* - h\frac{\sigma}{\sqrt{n}}]$$

où $n$ est la taille de l'échantillon et $h$ vérifie

$$\mathbb{P}(|U| \leq h) = 1 - \alpha \quad \text{et} \quad U \sim \mathcal{N}(0,\, 1)\,.$$

– Application numérique : pour $m^* = (x_1 + ...x_n)/n$, l'intervalle de confiance est

$$I = [m^* - h\frac{\sigma}{\sqrt{n}},\, m^* - h\frac{\sigma}{\sqrt{n}}]$$

**2.b. Estimation par intervalle de confiance de $m$ si $\sigma$ est inconnue.**

(sans détail)

Soit

$$T_n^* = \frac{M_n^* - m}{\sqrt{V_n^{*\prime}}/\sqrt{n}}$$

où $M_n^*$ est l'estimateur ponctuel de $m$ défini en 1.a et $V_n^{*\prime}$ est l'estimateur ponctuel de $\sigma^2$ défini en 1.c.

On peut montrer que la variable aléatoire $T_n^*$ suit une loi de Student à $n-1$ degrés de liberté.

De façon similaire au 2.a, on obtient :

*Estimation par intervalle de confiance de $m$ si $\sigma$ est inconnu*

– Seuil de confiance : $\alpha$ (ou confiance $1-\alpha$)

– Intervalle de confiance (aléatoire) :

$$I = [M_n^* - t_{n-1;\alpha}\frac{\sqrt{V_n^{*\prime}}}{\sqrt{n}},\, M_n^* + t_{n-1;\alpha}\frac{\sqrt{V_n^{*\prime}}}{\sqrt{n}}]$$

où $n$ est la taille de l'échantillon et $t_{n-1;\alpha}$ vérifie

$$\mathbb{P}(|T_n^*| \geq t_{n-1;\alpha}) = \alpha\,, \tag{3}$$

et $T_n^*$ suit une loi de Student à $n-1$ degrés de liberté.

– Application numérique :

Pour $m^* = (x_1 + ... + x_n)/n$ et $v_n^{*\prime} = \frac{1}{n-1}((x_1 - m^*)^2 + ... + (x_n - m^*)^2)$ l'intervalle de confiance est

$$I = [m^* - t_{n-1;\alpha}\frac{\sqrt{v_n^{*\prime}}}{\sqrt{n}},\, m^* + t_{n-1;\alpha}\frac{\sqrt{v_n^{*\prime}}}{\sqrt{n}}]$$

**2.c Estimation par intervalle de confiance de $\sigma^2$**

Soit la variable aléatoire

$$U_n^* = \frac{(n-1)V_n^{*\prime}}{\sigma^2}.$$

On peut montrer que $U_n^*$ suit une loi une loi du chi-deux à $n-1$ degrés de liberté.

On remarque que si $h_1$, $h_2$ dans $\mathbb{R}_+^*$

$$\{h_1 \leq U_n^* \leq h_2\} = \{h_1 \leq \frac{(n-1)V_n^{*\prime}}{\sigma^2} \leq h_2\} = \{\frac{(n-1)V_n^{*\prime}}{h_2} \leq \sigma^2 \leq \frac{(n-1)V_n^{*\prime}}{h_1}\}$$

On va donc chercher $h_1$ et $h_2$ tels que

$$\mathbb{P}(U_n^* < h_1) = \mathbb{P}(U_n^* > h_2) = \alpha/2\,, \tag{4}$$

ce qui impliquera $\mathbb{P}(h_1 \leq U_n^* \leq h_2) = 1-\alpha$. On notera alors $h_2 = \chi^2_{n-1;\alpha/2}$ et $h_1 = \chi^2_{n-1;1-\alpha/2}$.

*Estimation par intervalle de confiance de $\sigma^2$*

– Seuil de confiance $\alpha$ (ou confiance $1-\alpha$)

– Intervalle de confiance (aléatoire) :

$$I = [\frac{(n-1)V_n^{*\prime}}{\chi^2_{n-1;\alpha/2}},\ \frac{(n-1)V_n^{*\prime}}{\chi^2_{n-1;1-\alpha/2}}]$$

où $n$ est la taille de l'échantillon, et $\chi^2_{n-1;\alpha/2}$ et $\chi^2_{n-1;1-\alpha/2}$ vérifient $\mathbb{P}(U_n^* > \chi^2_{n-1;\alpha/2}) = \alpha/2$, et $\mathbb{P}(\chi^2_{n-1;1-\alpha/2}) \geq 1-\alpha/2$, et $U_n^*$ suit une loi du chi-deux à $n-1$ degrés de liberté.

– Application numérique : pour $v_n^{*\prime} = \frac{1}{n-1}((x_1 - m^*)^2 + ... + (x_n - m^*)^2)$ l'intervalle de confiance est

$$I = [\frac{(n-1)v_n^{*\prime}}{\chi^2_{n-1;\alpha/2}},\ \frac{(n-1)v_n^{*\prime}}{\chi^2_{n-1;1-\alpha/2}}]\,.$$

**2.d. Estimation par intervalle de confiance d'une proportion $p$**

*Estimation par intervalle de confiance d'une proportion p*

– Seuil de confiance $\alpha$ (ou confiance $1-\alpha$)

– Intervalle de confiance (aléatoire) :

$$I = [M_n^* - h\sqrt{\frac{M_n^*(1-M_n^*)}{n}},\ M_n^* + h\sqrt{\frac{M_n^*(1-M_n^*)}{n}}]$$

où $n$ est la taille de l'échantillon, $h$ vérifie

$$\mathbb{P}(|U| \leq h) = 1-\alpha \quad \text{et} \quad U \sim \mathcal{N}(0,\,1)\,.$$

– Application numérique :

$$I = [m_n^* - h\sqrt{\frac{m^*(1-m^*)}{n}},\ m_n^* + h\sqrt{\frac{m^*(1-m^*)}{n}}]\,.$$

II. TESTS D'HYPOTHÈSE

Dans un test, il ne s'agit plus de calculer ou d'estimer la valeur d'un paramètre, mais sur la foi de mesures effectuées, de décider d'accepter ou de rejeter une hypothèse $H_0$ posée a priori sur la loi $\mathcal{L}$ d'une variable aléatoire $X$ qui est liée au résultat d'une expérience $\mathcal{E}$.

On répète donc $n$ fois l'expérience $\mathcal{E}$ ce qui est modélisé par $n$ variables aléatoires $X_1$,..., $X_n$ indépendantes et de même loi que $X$.

On fait alors une hypothèse $H_0$ sur la loi $\mathcal{L}$, puis on se donne un **seuil de test** ou **risque** $\alpha > 0$ et un estimateur $Y = f(X_1, ..., X_n)$ du paramètre qui intervient dans l'hypothèse. On détermine ensuite un région $\mathcal{C} \subset \mathbb{R}$ appelée **région critique** telle que $\mathbb{P}(Y_n^* \in \mathcal{C}) \leq \alpha$ si l'hypothèse $H_0$ à tester est vraie.

Ainsi, si $H_0$ est vraie, il sera "peu probable" (proba égale à $\alpha$) de trouver une valeur de $Y_n^*$ dans la région critique. On concluera donc le test de la façon suivante, pour $X_1$,..., $x_n$ un échantillon de taille $n$ de $\mathcal{E}$ :

- Si $f(x_1, ..., x_n) \in \mathcal{C}$, on rejette l'hypothèse $H_0$ au risque $\alpha$
- Si $f(x_1, ..., x_n) \notin \mathcal{C}$, on accepte l'hypothèse $H_0$ au risque $\alpha$

Remarque : $\alpha$ représente la probabilité de rejeter l'hypothèse $H_0$ alors que celle-ci est vraie ; le réel $\alpha$ est alors appelé *risque de première espèce*. Il existe un autre risque, beaucoup moins utilisé, appelé risque de deuxième espèce, qui est la probabilité d'accepter l'hypothèse $H_0$ alors qu'elle est fausse.

**1. Tests sur les paramètres d'une loi normale.**

On suppose dans ce paragraphe que la loi de la variable aléatoire $X$ est la loi normale $\mathcal{N}(m,\,\sigma)$ où suivant les cas $m$ ou $\sigma$ sera la paramètre à tester.

Dans ce qui suit, on utilisera les mêmes notations que dans le paragraphe estimations. Ainsi, $M_n^*$ est défini par (1) et $V_n^{*\prime}$ par (2).

On notera aussi par $\pi_{1-\alpha/2}$ le réel tel que

$$\mathbb{P}(|U| \leq \pi_{1-\alpha/2}) = 1-\alpha \quad \text{où} \quad U \sim \mathcal{N}(0,\,1)\,.$$

*Test bilatéral sur la moyenne $m$ si $\sigma$ est connu.*
- Condition : $X$ est de loi $\mathcal{N}(m,\sigma)$ et $\sigma$ est supposé connu
- Hypothèse $H_0$ : "$m = m_0$"
- Risque $\alpha$
- Estimateur : $R_n^* = \frac{M_n^* - m_0}{\sigma/\sqrt{n}}$
- Région critique : $\mathcal{C} = \{\ r \in R;\ |r| > \pi_{1-\alpha/2}\} \ = \,]-\infty, -\pi_{1-\alpha/2}[\,\cup\,]\pi_{1-\alpha/2}, +\infty[$

Si on dispose d'une série statistique de valeurs $x_1$, $x_2$,..., $x_n$ pour $X$, on calcule alors la valeur associée pour $M_n^*$ puis pour $R_n^*$. Si cette dernière valeur est dans $\mathcal{C}$, on refuse l'hypothèse, sinon, on accepte l'hypothèse.

*Test unilatéral sur la moyenne $m$ si $\sigma$ est connu.*
- Condition : $X$ est de loi $\mathcal{N}(m,\sigma)$ et $\sigma$ est supposé connu
- Hypothèse $H_0$ : "$m \leq m_0$" (respectivement "$m \geq m_0$")
- Risque $\alpha$
- Estimateur : $R_n^* = \frac{M_n^* - m_0}{\sigma/\sqrt{n}}$
- Région critique : $\mathcal{C} = ]\pi_{1-\alpha/2}, +\infty[$ (respectivement $\mathcal{C} = ]-\infty,\ -\pi_{1-\alpha/2}[$).

*Test bilatéral sur la moyenne $m$ si $\sigma$ est inconnu.*
- Condition : $X$ est de loi $\mathcal{N}(m,\sigma)$ et $\sigma$ inconnu
- Hypothèse $H_0$ : "$m = m_0$"
- Risque $\alpha$
- Estimateur : $T_n^* = \frac{M_n^* - m_0}{\sqrt{V_n^{*\prime}}/\sqrt{n}}$
- Région critique : $\mathcal{C} = \{\ r \in R;\ |r| > t_{n-1,\alpha}\} \ = \,]-\infty, -t_{n-1,\alpha}[\,\cup\,]t_{n-1,\alpha},\ +\infty[$

où $t_{n-1,\alpha}$ est défini par (3).

*Test bilatéral sur l'écart type $\sigma$ si $m$ est inconnu.*
- Condition : $X$ est de loi $\mathcal{N}(m,\sigma)$ et $m$ inconnu
- Hypothèse $H_0$ : "$\sigma = \sigma_0$"
- Risque $\alpha$
- Estimateur : $U_n^* = \frac{(n-1)V_n^{*\prime}}{\sigma_0^2}$
- Région critique : $\mathcal{C} = \,]-\infty, -\chi_{n-1,1-\alpha/2}[\,\cup\,]\chi_{n-1,\alpha/2},\ +\infty[$.

où $\chi_{n-1,1-\alpha/2}$ et $\chi_{n-1,\alpha/2}$ sont définis par (4).

**2. Tests sur une proportion.**

*Test bilatéral sur une proportion.* Si $n \geq 50$ et $np^*(1-p^*) \geq 10$,
- Condition : $X$ est de loi de Bernoulli $\mathcal{B}(1,\,p)$.
- Hypothèse $H_0$ : "$p = p_0$"
- Risque $\alpha$
- Estimateur : $R_n^* = \frac{M_n^* - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$
- Région critique : $\mathcal{C} = \{\ r \in R;\ |r| > \pi_{1-\alpha/2}\} \ = \,]-\infty, -\pi_{1-\alpha/2}[\,\cup\,]\pi_{1-\alpha/2}, +\infty[$

**3. Test du $\chi^2$ ou test de conformité.**

Ce test permet de vérifier à partir d'observations si on peut raisonnablement considérer qu'une variable aléatoire $X$ suit une loi donnée, loi qui peut être maintenant quelconque.

Par exemple, on observe 200 familles de 3 enfants et l'on compte le nombre de filles de chaque famille. On trouve les résultats suivants :

| Nombre de filles | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| Nombre de familles | 20 | 83 | 70 | 27 |

On veut savoir si d'après ces résultats on peut considérer que la naissance d'une fille et la naissance d'un garçon sont des événements équiprobables ; c'est à dire si la loi de la variable aléatoire $X$ donnant le nombre de filles dans une famille de trois enfants est la loi $\mathcal{B}(3,\, 1/2)$.

*1er cas : $X$ prend un nombre fini de valeurs $a_1$,..., $a_q$.*

On teste dans ce cas l'hypothèse multiple $H_0$ : "$p_k = \mathbb{P}(X = a_k) = p_{0k}$, pour $k = 1, ..., q$" où les nombres $p_{0k} > 0$ sont donnés par la loi à laquelle on veut conformer la loi de $X$.

On réalise donc $n$ fois une expérience et pour $1 \leq k \leq q$, on appelle $O_k$ la variable aléatoire donnant le nombre de fois où $X$ prend la valeur $a_k$ ; $O_k$ est dit effectif observé pour la valeur $a_k$ et suit la loi binomiale $\mathcal{B}(n,\, p_k)$.

Si $H_0$ est vraie, l'espérance de $O_k$ : $np_k$ vaut $np_{0k}$ et $E_k = np_{0k}$ sera dit effectif espéré pour la valeur $a_k$.

Puis on considère la statistique $U_n^* = \sum_{k=1}^{q} \frac{(O_k - E_k)^2}{E_k}$ qui suit approximativement la loi $\chi^2_{q-1}$ (théorème de Cochran) si $H_0$ est vraie et si $n$ est assez grand pour que l'on puisse faire l'approximation de Moivre-Laplace et approcher la loi de $O_k$ par une loi normale. Dans la pratique, on exigera :

$$E_k \geq 5 \quad \text{pour} \quad k = 1, ...q\,.$$

Pour un risque $\alpha > 0$, on prendra pour région critique $\mathcal{C} = ]\chi_{q-1,\alpha},\, +\infty[$, région qui vérifie $\mathbb{P}(U_n^* \in \mathcal{C}) = \alpha$ et qui permet de rejeter les valeurs trop grandes de $U_n^*$, c'est à dire les écarts $O_k - E_k$ trop grands.

On résume à l'aide d'un tableau (et avec l'exemple donné ci-dessus)

| $a_k$ | $E_k$ | $O_k$ | $\dfrac{(O_k - E_k)^2}{E_k}$ |
|---|---|---|---|
| 0 | $\frac{1}{8} \times 200 = 25$ | 20 | 1 |
| 1 | $\frac{3}{8} \times 200 = 75$ | 83 | $0,853$ |
| 2 | $\frac{3}{8} \times 200 = 75$ | 70 | 1 |
| 3 | $\frac{1}{8} \times 200 = 25$ | 27 | $0,16$ |
| Total | 200 | 200 | $u^* = 3,013$ |

On a $q = 4$ ; d'après la table du chi-deux à $(q-1) = 3$ degrés de libertés, pour $\alpha = 0,05 = 5\%$, on a $\chi_{q-1,\alpha} = 7,82$. Donc $\mathcal{C} = ]\chi_{q-1,\alpha},\, +\infty[ = ]7,82,\, +\infty[$. Comme $u^* \notin \mathcal{C}$, on accepte l'hypothèse que la variable aléatoire qui donne le nombre de filles dans les familles de 3 enfants est donné par une loi binomiale $\mathcal{B}(3,\, 1/2)$.

*2ème cas : $X$ prend un nombre infini de valeurs.*

Dans ce cas là on se ramène au cas où $X$ prend un nombre fini de valeurs en consid{erant des intervalles au lieu de valeurs pour les $a_k$.

Les exercices

Estimations ponctuelles et estimations par intervalle de confiance

**Exercice 1.** Un échantillon de 30 cigarettes d'une même marque a donné les teneurs en goudron, exprimées en mg, suivantes :

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 12,9 | 12,7 | 12,4 | 12,8 | 14,5 | 13,1 | 12,9 | 14,5 | 11,7 | 12,3 |
| 13,4 | 12,8 | 13,4 | 12,9 | 12,9 | 12,5 | 12,5 | 12,8 | 11,8 | 11,8 |
| 13,4 | 12,4 | 13,4 | 12,5 | 12,8 | 12,8 | 13,5 | 12,8 | 12,9 | 12,7 |

La norme en vigueur recommande une teneur en goudron inférieure ou égale à 13 mg par cigarette.
i) Donnez une estimation ponctuelle de la proportion des cigarettes de cette marque qui respectent la norme de la teneur en goudron.
ii) Donnez une estimation ponctuelle de la teneur en goudron moyenne des cigarettes de cette marque, puis de l'écart type.
iii) On suppose maintenant que la variable aléatoire $X$ qui donne la teneur en goudron en mg d'une cigarette de cette marque suit une loi normale $\mathcal{N}(m, \sigma)$. Donnez avec un risque de 5% une estimation par intervalle de confiance de $m$. Peut-on en déduire, avec cette confiance, que la norme en vigueur sur la teneur en goudron est respectée ?

**Exercice 2.** Avant une élection, une société effectue un sondage sur le choix des votants entre deux candidats $A$ et $B$. Sur la population totale $F$ des votants, on interroge 1204 personnes, et on obtient les intentions de vote suivantes :

$$662 \text{ pour } A \quad \text{et} \quad 542 \text{ pour } B$$

A l'aide de ces résultats, donner, pour une personne quelconque de la population totale $F$, une estimation par intervalle de confiance de l'intention de vote pour le candidat $A$, avec un risque de 10%, puis un risque de 5%.

**Exercice 3.** On mesure la force de compression d'un ciment en moulant de petits cylindres et en mesurant la pression $X$, exprimée en kg.cm$^{-2}$, à partir de laquelle ils se brisent. A partir des mesures sur un échantillon de 10 cylindres, on a calculé une moyenne statistique $\overline{x} = 19,72$, et une variance (non corrigée) $\overline{v} = 0,6096$.
i) Donner une estimation (ponctuelle) pour $\mathbb{E}(X)$ et $\mathrm{Var}(X)$.
ii) En supposant $X$ gaussien, donner un intervalle de confiance avec un seuil de confiance $\alpha = 0,1$, de $\mathbb{E}(X)$ et $\mathrm{Var}(X)$.
iii) On suppose maintenant que la variance $\mathrm{Var}(X)$ vaut $0,69$. Donner un nouvel intervalle de confiance pour l'espérance de $X$ avec le risque $0,1$.

**Exercice 4.** Un laboratoire d'agronomie a effectué une étude sur le maintien du pouvoir germinatif des graines de *Papivorus subaquaticus* après une conservation de 3 ans. Sur un lot de 80 graines, 47 ont germé. Donner, avec une confiance de 95%, un intervalle de confiance pour la probabilité de germination des graines de Papivorus subaquaticus après une conservation de 3 ans.

**Exercice 5.** On a étudié la répartition du poids de 100 jeunes gens et on a obtenu le tableau suivant :

| Poids en kg | $[40,45[$ | $[45,50[$ | $[50,55[$ | $[55,60[$ | $[60,65[$ | $[65,70[$ | $[70,75[$ |
|---|---|---|---|---|---|---|---|
| Fréquence | 5 | 12 | 31 | 31 | 16 | 3 | 2 |

Estimer les paramètres de la loi normale qui modélise ces observations.

**Exercice 6.** Un laboratoire indépendant a vérifié pour le compte de l'office de la protection du consommateur, la résistance à l'éclatement (en kg.cm$^{-2}$) d'un réservoir à essence d'un certain fabricant. Des essais similaires, effectués il y a un an, permettent de considérer que la résistance à l'éclatement est distribuée normalement avec une variance de 9. Des essais sur un échantillon de 10 réservoirs conduisent aux mesures de résistance suivantes (en kg.cm$^{-2}$) :

$$211;\ 234,5;\ 213,5;\ 228,5;\ 225;\ 219;\ 207;\ 241;\ 212,5;\ 198\ .$$

Estimer par un intervalle de confiance la résistance moyenne à l'éclatement de ce type de réservoir avec une confiance de 95%, puis de 99%.

**Exercice 7.** Un sondage sur la popularité du premier ministre indique que 51% des personnes interrogées sont favorables à sa politique. Construire un intervalle de confiance de niveau $0,95$ pour la proportion $p$ de français favorables à cette politique sachant que ce sondage a été réalisé auprès de $n = 100$ personnes ; même question si $n = 1000$. Quelle aurait du être la taille de l'échantillon pour que l'intervalle soit de longueur inférieure à 4% ?

Tests d'hypothèses

**Exercice 8.** On souhaite contrôler un lot important de pièces métalliques. Une caractéristique des ces pièces, notée $X$, est une variable aléatoire, dont la loi de probabilité est normale de moyenne $m$ et de variance $\sigma^2$ inconnues. On se propose de prendre une décision au sujet de $\sigma^2$ au vu d'un échantillon de taille $n = 12$, $\{x_1, x_2, \ldots, x_{12}\}$. On sait également que

$$\sum_{i=1}^{12} \left( x_i - \frac{\sum_{i=1}^{12} x_i}{12} \right)^2 = 650\ .$$

Doit-on accepter l'hypothèse "$\sigma^2 = 100$" avec un risque de $0,05$ ?

**Exercice 9.** On se demande quelle est la proportion $p$ des ménages possédant un poste de télévision. Pour déterminer le paramètre $p$ on hésite entre deux hypothèses $H_0$ et $H_1$. On se propose de prendre une décision au vu d'un échantillon significatif de taille $n = 125$. Soit $f_n$ la proportion observée dans l'échantillon. On obtient $f_n = 0,48$. On prendra un risque de $0,05$. Doit-on accepter l'hypothèse $H_0$ : "$p = 0,5$" ?

**Exercice 10.** Etant donné un certain dé on veut tester l'hypothèse $H_0$ : "Le dé est régulier".

On note note $X$ la variable aléatoire définie par le chiffre obtenu si on lance le dé

i) Déterminer $\mathbb{P}(X = 1),\ \mathbb{P}(X = 2), \ldots,\ \mathbb{P}(X = 6)$.

ii) On lance 600 fois le dé. Quels sont les effectifs $E_1$ (respectivement $E_2, \ldots E_6$), effectifs espérés pour "$X = 1$" (respectivement "$X = 2$",... ,"$X = 6$") ?

iii) On a obtenu 105 fois le chiffre 1, 98 fois le chiffre 2, 103 fois le chiffre 3, 111 fois le chiffre 4, 95 fois le chiffre 5 et 88 fois le chiffre 6. Peut-on accepter l'hypothèse $H_0$ avec le risque $0,05$ ?

**Exercice 11.** On lance 4000 fois une pièce de monnaie. On obtient 1870 faces. La pièce est-elle truquée ?

**Exercice 12.** On suppose que le taux de chômage dans un pays donné est de 10%. Dans une ville donnée, on observe 1080 chômeurs sur 10000 personnes actives. Peut-on dire que la différence avec le pourcentage national est significatif, avec un risque de 5% ?

**Exercice 13.** Le prix d'un même article relevé au hasard dans 15 épiceries de la ville donne ceci :

$$\begin{array}{cccccccc} 42,7 & 42,6 & 43 & 43,5 & 42,8 & 43,1 & 43,6 & 42,9 \\ 41,6 & 42,8 & 42,9 & 43,2 & 42,6 & 43,1 & 43,1 & \end{array}$$

On admet que $X$ est une variable aléatoire suivant une loi gaussienne. On fixe les risques à 5%.
i) Peut-on admettre l'hypothèse $\mathbb{E}(X) = 43,0$ ?
ii) Peut-on admettre l'hypothèse $\mathrm{Var}(X) = 0,1$ ?

**Exercice 14.** Pour justifier la loi de Gompertz modélisant l'extinction des populations, K. Miescher a mesuré l'évolution au cours du temps du nombre de rats vivants dans un élevage qu'il commença avec 144 rats en 1953. Le tableau suivant donne le nombre $N$ de rats vivants au bout de $t$ mois.

| $t$ | 10 | 15 | 20 | 25 | 28 | 30 | 32 | 34 | 36 | 38 | 40 | 42 | 43 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $N$ | 144 | 143 | 140 | 131 | 119 | 106 | 86 | 63 | 37 | 15 | 4 | 1 | 0 |

i) Remplir le tableau suivant donnant l'effectif $E$ des rats ayant une durée de vie $D$ dans chaque intervalle de temps.

| $D$ | ]10, 20] | ]20, 25] | ]25, 30] | ]30, 34] | ]34, 38] | ]38, 40] | ]40, 43] |
|---|---|---|---|---|---|---|---|
| $E$ | | | | | | | |

ii) Tracer l'histogramme correspondant. L'allure de l'histogramme ressemble-t-il à celui que donnerait une loi gaussienne ?
iii) Evaluer la moyenne $m$ et l'écart type $\sigma$ de la variable aléatoire $D$ en supposant que sa loi est Gaussienne.
iv) A l'aide d'un test du $\chi^2$, vérifier que la différence entre les résultats expérimentaux et les prédictions données par la loi Gaussienne (dont on fixe les paramètres $m$ et $\sigma$ conformément à l'estimation précédente) est significative avec un seuil élevé.

## Quelques exercices corrigés

**Exercice**[Estimations ponctuelles d'une espérance et d'une variance - estimation par intervalles de confiance d'une moyenne (cas variance inconnue et cas variance connue)]

Chez 10 patients qui ne pratiquent aucune activité sportive, on fait un relevé du taux de glycémie (en $g/\ell$). On obtient les résultats suivants :

| 0.96 | 1.2 | 0.80 | 1.08 | 1 | 1.05 | 1.11 | 1.3 | 1.05 | 1.34 |
|---|---|---|---|---|---|---|---|---|---|

**i)** Justifier (empiriquement) pourquoi il est raisonnable de supposer que la variable aléatoire $X$ qui décrit le taux de glycémie d'une population donnée suit une loi normale. Dans toute la suite de cet exercice, on supposera $X \sim \mathcal{N}(m, \sigma)$.
**ii)** Donner une estimation ponctuelle de la moyenne $m$.
**iii)** Donner une estimation ponctuelle de la variance $\sigma^2$.
**iv)** Dans une population usuelle, le taux de glycémie suit une variable aléatoire normale $\mathcal{N}(m, \sigma)$ de paramètres $m = 1$ et $\sigma = 0,1$. Avec un risque de 5%, dire si on peut considérer que le taux de glycémie moyen de ces 10 patients est conforme à celui d'une population usuelle (i.e., appartient aux intervalles de confiance) :
**a)** Dans le cas où on suppose l'écart type connu et égal à $0,1$.
**b)** Dans le cas où on suppose l'écart type inconnu.

**Solution.**

**i)** Dans la mesure où la taille $N$ de population étudiée est "grande", il est naturel de supposer que $X$ suit une loi normale, en utilisant le théorème de la limite centrale, puisque $X$ apparaît comme somme de $N$ variables aléatoires (quantitatives) indépendantes et de même loi.
**ii)** Une estimation ponctuelle $m^*$ de l'espérance de $X$ est

$$m^* = \frac{0.96 + 1.2 + ... + 1.34}{10} = 1.089 \ .$$

**iii)** Une estimation ponctuelle $v^{*\prime}$ de la variance est

$$v^{*\prime} = \frac{(0.96 - m^*)^2 + ... + (1.34 - m^*)^2}{10 - 1} \approx 0.026.$$

**iv)** On a $n = 10$, $\alpha = 0.05$.

a) Dans le cas où $\sigma$ est supposé connu et égal à 0.1, l'intervalle de confiance $I$ pour la moyenne est donné par

$$I = [m^* - h\frac{\sigma}{\sqrt{n}},\ m^* + h\frac{\sigma}{\sqrt{n}}] \ ,$$

où $h$ est déterminé par $\Phi(h) = 1 - \alpha/2 = 0.975$. Cela donne $h = 1.96$ et donc

$$I = [1.027,\ 1.151] \ .$$

b) Dans le cas où $\sigma$ est supposé inconnu, l'intervalle de confiance $I$ est donné par

$$I = [m^* - h\frac{\sqrt{v^{*\prime}}}{\sqrt{n}},\ m^* + h\frac{\sqrt{v^{*\prime}}}{\sqrt{n}}] \ ,$$

où $h$ est déterminé par $\mathbb{P}(|T^*_{n-1}| > h) = \alpha$, et $T^*_{n-1} = T^*_9$ est une variable aléatoire qui suit une loi de Student de paramètre 9. Cela donne $h = 2.262$ et donc

$$I = [0.973,\ 1.204] \ .$$

**Exercice**[Estimation ponctuelle et par intervalle de confiance d'une proportion]

Un sondage effectué avant les présidentielles de mai 2007 auprès de 954 personnes, prédisait le résultat suivant pour le second tour des élections :

- S. Royal : 48%
- N. Sarkozy : 52%

**i)** A l'aide du sondage mentionné ci-dessus, donnez une estimation du pourcentage de votes pour les deux candidats du second tour.
**ii)** Donnez une estimation par intervalle de confiance du pourcentage de votes pour le candidat N. Sarkozy, avec une confiance de 90%, puis avec une confiance de 80%.
**iii)** Le 6 mai 2007, le résultat officiel du vote était :

- S. Royal : $46,94\%$
- N. Sarkozy : $53,06\%$.

En utilisant le résultat de la question ii), commentez le résultat du sondage.
iv) En supposant que quelque soit le nombre $N$ de personnes interrogées, le résultat du sondage aurait été le même ($52\% - 48\%$), donnez une estimation du nombre minimum $N$ de personnes qu'il aurait fallu interroger pour être sûr à 95% d'annoncer la victoire du candidat N. Sarkozy.

**Solution.**
**i)** Une estimation $p^*$ du pourcentage $p$ de votes au second tour pour le candidat Sarkozy est $p^* = 54\%$.

Une estimation du pourcentage de votes pour la candidate Royal est $p^* = 46\%$.
**ii)** a) Avec confiance 90% :

On a $n = 954$, $\alpha = 10\%$. Une estimation par intervalle de confiance de la proportion $p$ est

$$I = \left[p^* - h\frac{\sqrt{p^*(1-p^*)}}{\sqrt{n}},\ p^* + h\frac{\sqrt{p^*(1-p^*)}}{\sqrt{n}}\right] = [49.33\%,\ 54.67\%]$$

où $h$ a été déterminé par l'équation $\Phi(h) = 1 - \alpha/2 = 0.95$, i.e. $h \approx 1.65$.

b) Avec confiance 80% :

On a $n = 954$, $\alpha = 20\%$. On obtient dans ce cas $h \approx 1.29$ et

$$I = [49.91\%,\ 54.04\%] \ .$$

**iii)** Dans les deux cas précédents, la valeur réelle de $p$ est bien dans l'intervalle de confiance. De ce point de vue, le sondage était correct.

**iv)** Après le sondage, pour pouvoir être sûr (à 95%) d'annoncer la victoire du candidat Sarkozy, il aurait été nécessaire que la borne inférieure de l'intervalle de confiance pour $p$ soit au dessus de 50%. C'est à dire, avoir une valeur de $N$ au moins égale à celle qui résoud

$$p^* - h\frac{\sqrt{p^*(1-p^*)}}{\sqrt{N}} \geq 50\% \ ,$$

où $h$ vérifie $\Phi(h) = 1 - \alpha/2 = 0.975$, i.e., $h \approx 1.96$.

On trouve $N \geq 2397.16$, c'est à dire $N = 2398$.

**Exercice**[Test sur une proportion] Une roulette possède 37 numéros (de 0 à 36). Sur 10000 parties, le zéro est sorti 298 fois. On admet que la sortie du zéro est une variable aléatoire de Bernouilli.

On considère l'hypothèse simple le zéro sort avec une probabilité de 1/37. Avec un risque de 5%, peut-on accepter l'hypothèse ?

**Correction** A partir des résultats statistiques, donnons un intervalle de confiance de la probabilité pour que le zéro sorte (estimation d'une proportion).

L'estimation pontuelle donne, $p^* = 298/10000 \approx 2,98\%$. Avec un risque $\alpha = 0,05 = 5\%$, puisque le nombre de tirages est suffisamment élevé, on peut faire l'approximation de Moivre-Laplace, i.e., remplacer la loi binomiale $\mathcal{B}(n = 10000, p = 1/37)$ par la loi normale $\mathcal{N}(np, \sqrt{np(1-p)})$ ; En effet, on a $n = 10000 \geq 50$ et $np^*(1-p^*) \approx 289 \geq 10$.

On a donc comme intervalle de confiance de la probabilité (ou proportion) que le zéro sorte :

$$I = \left[p^* - h\sqrt{\frac{p^*(1-p^*)}{n}},\ p^* + h\sqrt{\frac{p^*(1-p^*)}{n}}\right] \ ,$$

où $h = \Phi^{-1}(1 - \alpha/2)$ et $\Phi$ est la fonction de répartition de la loi normale centrée réduite $\mathcal{N}(0,1)$. Pour $p^* \approx 0,0298$, $n = 10000$ et $\alpha = 0,05$ (qui donne $h = 1,96$) on trouve

$$I \approx [2,65\%,\ 3,31\%] \ .$$

La probabilité théorique, si le jeu n'est pas truqué, est de $1/37 \approx 2,70\% \in I$. Donc, on accepte l'hypothèse.

**Exercice**[Test du khi-deux] Selon la théorie Mendélienne de l'hérédité, on prévoit qu'en croisant deux types de plantes, on doit obtenir des variétés de quatre types (disons $S$, $T$, $U$ et $V$), avec une probabilité respective de 9/16, 3/16, 3/16, 1/16. A la suite d'une expérience, on obtient 154 plantes de type $S$, 44 de type $T$, 63 de type $U$ et 21 de type $V$.

Peut-on accepter l'hypotèse que l'échantillon obtenu est conforme à la théorie Mendélienne avec un risque de 5% ?

**Correction.** Il s'agit ici de tester une hypothèse multiple. L'effectif total de l'échantillon statistique est de $n = 154 + 44 + 63 + 21 = 282$. Avec les notations usuelles, on obtient le tableau suivant, pour les probabilités théoriques $p_{01} = \mathbb{P}(S) = 9/16$, $p_{02} = \mathbb{P}(T) = 3/16$, $p_{03} = \mathbb{P}(U) = 3/16$ et $p_{04} = \mathbb{P}(V) = 1/16$ :

| $a_k$ | $E_k = np_{0k}$ | $O_k$ | $(O_k - E_k)^2/E_k$ |
|---|---|---|---|
| $S$ | $158,625$ | 154 | $0,135$ |
| $T$ | $52,875$ | 44 | $1,490$ |
| $U$ | $52,875$ | 63 | $1,939$ |
| $V$ | $17,625$ | 21 | $0,646$ |
| Totaux | 282 | 282 | $u^* := 4,21$ |

Puisque pour tout $k = 1, 2, 3, 4$ on a $E_k \geq 5$, on peut utiliser l'approximation gaussienne et faire un test du chi-deux pour cette hypothèse multiple. La région critique est

$$\mathcal{C} = ]\chi^2_{q-1;\alpha};\ +\infty[\ ,$$

où $q = 4$ est le nombre d'hyptohèses simples, et $\alpha = 0,05$ est le risque. On obtient, par la lecture de la table de la loi du chi-deux :

$$\mathcal{C} = ]7,81, +\infty[\ .$$

Comme $u^* \notin \mathcal{C}$, on accepte l'hypothèse de conformité.

**Exercice**[Estimations] L'article suivant est extrait du journal **Le Monde** du *3 mars 1983*.

BOURSE DE NEW-YORK
NOUVEAU RECORD

La reprise frappe à la porte. Wall Street en est maintenant convaincu, après la publication faite par le Département du Commerce des principaux indicateurs économiques pour janvier.

Dans ces statistiques, il ressort que l'indice des valeurs industrielles a monté de 3,6%. Cette hausse mensuelle est la plus forte enregistrée depuis 1950. Elle est surtout supérieure aux prévisions les plus optimistes que les boursiers avaient pu faire.

Beaucoup sont maintenant persuadés autour du "Big board" que le marché est entré dans une nouvelle phase d'ascension. La clientèle particulière a, pour sa part, fait un retour très marqué que certains jugent significatif.

Sur les 1970 valeurs traitées le 2 mars, 1168 ont monté, 469 ont baissé et 333 n'ont pas varié.

Voici une sélection des cours du jour :

| *Valeurs* | *Cours du* 1er *mars* | *Cours du* 2 *mars* |
|---|---|---|
| Alcoa | 34 3/4 | 35 1/8 |
| ATT | 67 1/2 | 66 7/8 |
| Boeing | 37 | 36 7/8 |
| Chase Manhattan Bank | 47 1/4 | 48 7/8 |
| Du Pont de Nemours | 40 5/8 | 41 |
| Eastman Kodak | 89 1/4 | 89 |
| Exxon | 30 | 30 7/8 |
| Ford | 40 1/2 | 41 3/8 |
| General Electric | 111 | 108 |
| General Foods | 39 3/8 | 39 1/2 |
| General Motors | 63 1/2 | 63 |
| Goodyear | 31 3/4 | 31 5/8 |
| IBM | 101 3/4 | 102 1/8 |
| ITT | 33 3/8 | 33 3/4 |
| Mobil Oil | 27 1/4 | 28 1/2 |
| Pfizer | 72 3/4 | 74 1/4 |
| Schlumberger | 40 1/2 | 41 7/8 |
| Texaco | 32 1/8 | 33 |
| U.A.L. Inc. | 34 | 34 3/8 |
| Union Carbide | 61 1/2 | 61 1/4 |
| U.S. Steel | 22 3/4 | 22 7/8 |
| Westinghouse | 48 5/8 | 49 1/4 |
| Xerox Corp | 38 5/8 | 39 5/8 |

i) D'après la sélection des valeurs dont les cours ont été publiés, donner un intervalle de confiance pour la proportion de valeurs ayant strictement monté lors de la séance du 2 mars. Que penser du résultat ?
ii) D'après les mêmes données, quel est l'intervalle de confiance pour la proportion de valeurs ayant vu leur cours inchangé ? Quelles reflexions supplémentaires ce résultat inspire-t-il ?

**Correction**

i) Le nombre de valeurs données est 23. Le nombre de valeurs ayant augmenté est de 16. L'estimation ponctuelle de la proportion est donc $p^* = 16/23 \approx 0,696$.

L'estimation par intervalle de confiance pour une proportion donne alors, pour un risque $\alpha$

$$I = \left[ p^* - h\sqrt{\frac{p^*(1-p^*)}{23}},\ p^* + h\sqrt{\frac{p^*(1-p^*)}{23}} \right]$$

Où $h = \Phi^{-1}(1-\alpha/2)$ ($\Phi$ est la fonction de répartition de la loi normale centrée réduite).

L'application numérique avec $\alpha = 0,05$ (choix canonique lorsqu'on n'impose pas de valeur) et la lecture de la table de la loi normale, donne $h = 1,96$ et donc

$$I = [50,1\%,\ 88,4\%]\ .$$

On en déduit que le choix de la sélection des cours du jour est représentative, car la proportion réelle des valeurs ayant augmenté est de $1168/1970 \approx 59,3\% \in I$.
ii) L'estimation pontuelle de la proportion des valeurs ayant vu leur cours inchangé est $q^* = 0/23 = 0$. On ne peut pas donner d'intervalle de confiance pour cette proportion, et bien évidemment, quel que soit le critère choisi, cette estimation ponctuelle de la vraie valeur $q = 333/1970 \approx 16,9\%$, est mauvaise.