UE Probabilité et Statistique - Mathématiques

Devoir Maison n°8

Vincent Escoffier, Adrien Jallais, Théo Martel, Louis Muzellec.

17 May 2022

Contents

Préparation de l'environnement

R et Rstudio seront utilisés pour la rédaction de ce DM en Rmarkdown, ainsi que les packages suivants :

```
library(readr)
library(dplyr)
library(ggplot2)
library(knitr)
library(ggpubr)
```

Exercice 1

Contexte

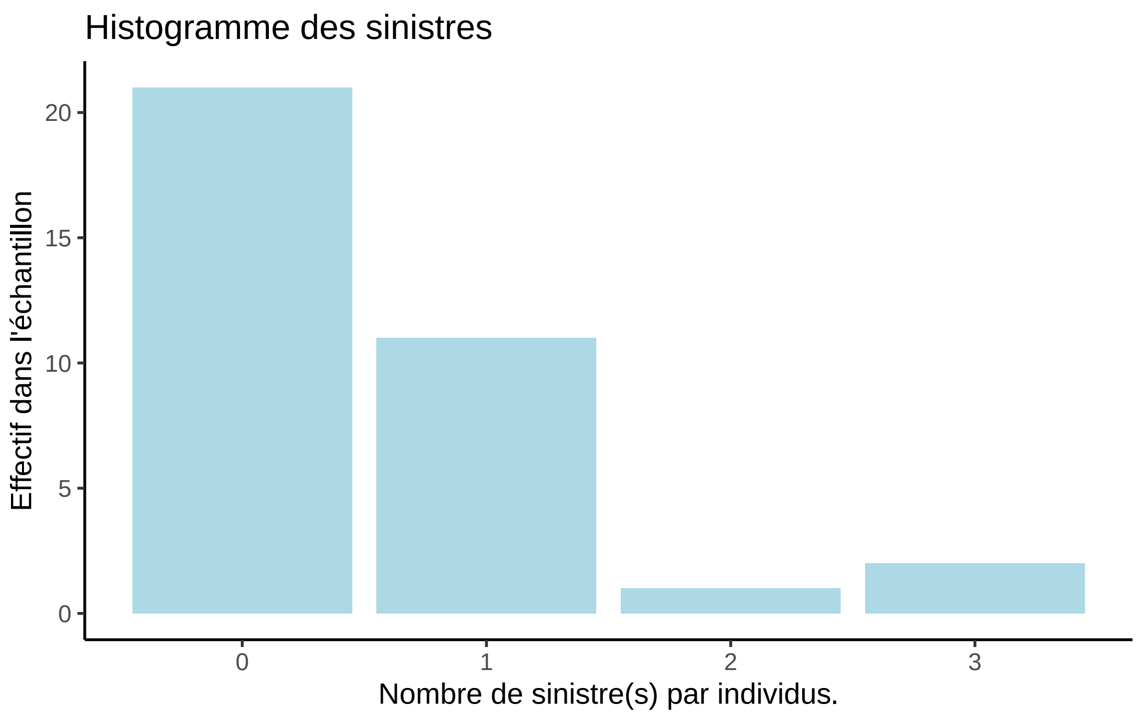
Soit X, la variable aléatoire représentant le nombre d'accidents par assuré. X est une variables aléatoire discrète et il est admis que l'occurence des sinistres X suit une loi de Poisson pour laquelle on recherche le paramètre inconnu $\lambda > 0$ (notée $\mathcal{P}_\lambda$).

Pour cela, on a disposition un échantillon aléatoire bernouillien $\underline{X}_n = (X_1, \dots, X_n)$, $n \geq 1$. Le modèle d'échantillonage associé est le suivant : $(\mathbb{N}, (\mathcal{P}_\lambda)_{\lambda>0})^n$.

La statistique de test associée à cet échantillon est $n\overline{X}_n$, qui représente le nombre total d'observations.

Description des données

L'histogramme suivant résument la dispersion de nos données :



L'effectif de l'échantillon est de 35. A partir de $\overline{X}_n = \frac{\sum_{i=0}^n x_i}{n} = 0.54$, on a $n\overline{X}_n = 19$.

Intervalle de confiance (IC) unilatéral pour λ

Pour rechercher une estimation de λ, on peut utiliser, à distance finie, la fonction pivotale $n\overline{X}_n$ qui suit une loi de Poisson de paramètre $n\lambda$: $n\overline{X}_n \rightsquigarrow \mathcal{P}_{n\lambda}$.

Notre modèle étant discret, avec le livret du module 2 du cours (page 45), on peut déduire l'intervalle de confiance par excès du paramètre λ :

$$IC_{\underline{X}_n}^{1-\alpha+}(\lambda) = \left[\frac{\chi^2_{2n\overline{X}_n}(\alpha_2)}{2n} ; \frac{\chi^2_{2(n\overline{X}_n+1)}(1-\alpha_1)}{2n} \right] \text{ avec } \alpha = \alpha_1 + \alpha_2 : (IC_1)$$

Or la table statistique des fractiles de la loi du χ^2 fournie, ne nous permet pas de connaître les fractiles d'un χ^2 de degré de liberté supérieur à 30. En effet, on a $2n\overline{X}_n = 38$ et $2(n\overline{X}_n + 1) = 40$. Nous ne pouvons donc utiliser l'IC_1.

Ainsi on considèrera l'effectif de notre échantillon grand. Dès lors, on donne la fonction pivotale asymptotique suivante :

$$\frac{\sqrt{n}(\overline{X}_n - \lambda)}{\sqrt{\overline{X}_n}} \xrightarrow{CL} \mathcal{N}(0,1)$$

On en déduit donc un intervalle de confiance par excès (IC_2) asymptotique, en utilisant les fractiles de loi normale :

$$IC_{\underline{X}_n}^{1-\alpha,\text{ asymp}}(\lambda) = \left[\overline{X}_n - Z_{1-\frac{\alpha}{2}} \times \sqrt{\frac{\overline{X}_n}{n}}; \overline{X}_n + Z_{1-\frac{\alpha}{2}} \times \sqrt{\frac{\overline{X}_n}{n}}\right] : (IC_2)$$

Dans la suite de cet exercice, nous utiliserons l'IC_2 pour l'estimation du paramètre λ.

1. Calcul d'un interval de confiance asymptotique unilatéral gauche de niveau 95% pour λ (IC_g)

On obtient un IC_g tel que :

$$IC_{\underline{X}_n}^{1-\alpha,\text{ asymp}}(\lambda) = \left[\overline{X}_n - Z_{1-\frac{\alpha}{2}} \times \sqrt{\frac{\overline{X}_n}{n}} < \lambda\right] : (1)$$

$$(1) \Leftrightarrow IC_{\underline{X}_n}^{0{,}95,\text{ asymp}}(\lambda) = \left[0,543 - Z_{1-\frac{0{,}05}{2}} \times \sqrt{\frac{0,543}{35}} < \lambda\right]$$

$$(1) \Leftrightarrow IC_{\underline{X}_n}^{0{,}95,\text{ asymp}}(\lambda) = \left[0,543 - 1,96 \times \sqrt{\frac{0,543}{35}} < \lambda\right]$$

$$(1) \Leftrightarrow IC_{\underline{X}_n}^{0{,}95,\text{ asymp}}(\lambda) = [0,298 < \lambda]$$

2. Calcul d'un interval de confiance asymptotique unilatéral droit de niveau 95% pour λ (IC_d)

On obtient un IC_d tel que :

$$IC_{\underline{X}_n}^{1+\alpha,\text{ asymp}}(\lambda) = \left[\overline{X}_n + Z_{1-\frac{\alpha}{2}} \times \sqrt{\frac{\overline{X}_n}{n}} > \lambda\right] : (2)$$

$$(2) \Leftrightarrow IC_{\underline{X}_n}^{0{,}95,\text{ asymp}}(\lambda) = \left[0,543 + Z_{1-\frac{0{,}05}{2}} \times \sqrt{\frac{0,543}{35}} > \lambda\right]$$

$$(2) \Leftrightarrow IC_{\underline{X}_n}^{0{,}95,\text{ asymp}}(\lambda) = \left[0,543 + 1,96 \times \sqrt{\frac{0,543}{35}} > \lambda\right]$$

$$(2) \Leftrightarrow IC_{\underline{X}_n}^{0{,}95,\text{ asymp}}(\lambda) = [0,787 > \lambda]$$

Commentaires

L'utilisation d'intervalles de confiance unilatéraux est inadapté pour les lois symétriques comme la loi normale, si nous en avions eu la possibilité, il aurait fallu utiliser l'IC_1.

Peut-on accepter l'hypothèse que $\lambda = 1$?

Nous répondrons à cette question, en évaluant un test d'hypothèse.

Test d'hypothèses

On pose $H_0 : \lambda = 1$ contre $H_1 : \lambda \neq 1$.

Observations

On observe que selon H_0, on a $\lambda \notin IC_g$ et $\lambda \notin IC_d$. Autrement dit $\lambda \notin [0.298; 0.787]$

Décision

On rejette H_0, jusqu'à preuve du contraire.

Conclusion

Par conséquent, on ne peut accepter l'hypothèse $\lambda = 1$, et donc on a $\lambda \neq 1$.

Commentaires

On peut vérifer cette négation, avec l'estimateur $W^{-\alpha}_{\underline{X}_n}(\lambda)$ de la manière suivante :

$$W^{\alpha}_{\underline{X}_n}(\lambda) = \left[\frac{\underline{X}_n}{\overline{X}_n} < a_\alpha\right] \text{ ou } W^{\alpha}_{\underline{X}_n}(\lambda) = \left[\overline{X}_n > b_\alpha\right]$$

$$P_{\lambda=1}(W^{-\alpha}_{\underline{X}_n}(\lambda)) \leq \alpha \text{ et } P_{\lambda\neq1}(W^{\alpha}_{\underline{X}_n}(\lambda)) \geq \alpha$$

Exercice 2

Contexte

Soit X, une variable aléatoire discrète représentant le nombre de pièces défectueuses par échantillon (représentée par une commande ou un lot). Les pièces peuvent être soient défectueuses, soient fonctionnelles. Notre échantillonage est donc extrait d'une épreuve de Bernouilli. La taille de l'échantillon est grande : $n > 100$ avec $n = 140$.

On souhaite savoir si le client acceptera cet echantillon, et pour cela il faut qu'il contienne au moins 120 composants fonctionnels.

Description du modèle de données

Notre échantillonnage aléatoire est simple tel que : $X_n = (X_i)_{1\leq i\leq n}$ avec le modèle suivant : $(0,1(\mathcal{B}(1,p)_p)_{p\in[0,1]})$ sachant $p = 10\% = 0,1$.

On va comparer ce paramètre avec une estimation ponctuelle, afin de savoir si l'affirmation du fabricant est vraie. Pour cela nous allons réaliser deux tests : le premier du point de vue du fabricant (TH_1), le second du point de vue du client (TH_2).

Observation ponctuelle

Dans notre observation ponctuelle (lot), on a observé que la proportion de pièces défectueuses est de 0.0857 ($= 12/140$).

Point de vue du fabricant (TH_1)

Test d'hypothèses

$$H_0 : p_0 \leq 0,1 \text{ contre } H_1 : p_0 > 0,1$$

$$W_{X_n}^{\alpha,\text{ asymp}}(p) = \left\{ \frac{X_n}{\overline{\overline{X}}_n} > p + \frac{Z_{1-\alpha}}{\sqrt{n}} \times \sqrt{p(1-p)} \right\} : (3)$$

$$(3) \Leftrightarrow W_{140}^{0{,}05,\text{ asymp}}(p) = \left\{ \frac{X_n}{\overline{\overline{X}}_n} > 0,1 + \frac{1,96}{\sqrt{140}} \times \sqrt{0,1(1-0,1)} \right\}$$

$$(3) \Leftrightarrow W_{140}^{0{,}05,\text{ asymp}}(p) = \left\{ \frac{X_n}{\overline{\overline{X}}_n} > 0,1496 \right\}$$

Observations

$$\frac{X_n}{\overline{\overline{X}}_n} = 0,0857 \notin W_{X_n}^{\alpha,\text{ asymp}}(p)$$

Décision

Pour le TH_1, on ne rejette pas l'hypothèse H_0 jusqu'à preuve du contraire.

Conlcusion

Jusqu'à preuve du contraire, le client acceptera le lot.

Point de vue du client (TH_2)

Test d'hypothèses

$$H_0 : p_0 \geq 0,1 \text{ contre } H_1 : p_0 < 0,1$$

$$W_{X_n}^{\alpha,\text{ asymp}}(p) = \left\{ \frac{X_n}{\overline{X}_n} > p - \frac{Z_{1-\alpha}}{\sqrt{n}} \times \sqrt{p(1-p)} \right\} : (3)$$

$$(3) \Leftrightarrow W_{140}^{0{,}05,\text{ asymp}}(p) = \left\{ \frac{X_n}{\overline{X}_n} > 0,1 - \frac{1,96}{\sqrt{140}} \times \sqrt{0,1(1-0,1)} \right\}$$

$$(3) \Leftrightarrow W_{140}^{0{,}05,\text{ asymp}}(p) = \left\{ \frac{X_n}{\overline{X}_n} > 0,0503 \right\}$$

Observations

$$\frac{X_n}{\overline{X}_n} = 0,0857 \notin W_{X_n}^{\alpha,\text{ asymp}}(p)$$

Décision

Pour le TH_2, on accepte donc H_0.

Conlcusion

Jusqu'à preuve du contraire, le client n'acceptera pas le lot. **??Problème de formulation ou de conclusion -> à vérifier??**

Commentaires

L'acceptation du lot a été évaluée depuis les points de vues des deux parties.

On se place du point de vue du client, il ne faudrait donc pas tester TH_2 mais seulement TH_1. En effet, en se placant du côté du client, on souhaite seulement évaluer ce que craint le client de rejeter à tord. On a réalisé deux tests, si ilssont contradictaatoires => prendre le test du point de vue client.

??? TODO : vérifier la formulation des deux tests, avec le livret de cours module 2 (page 62).????

Exercice 3

Contexte

Une expérience chimique consiste à ajouter une dose x d'un agent A dans une solution S. Après réaction, on mesure la quantité d'une espèce E.

Pour x donné, on suppose qu'il est pertinent de représenter cette mesure par une variable aléatoire $Y = \theta x^2 + aZ$, a ($a > 0$) connu, θ ($\theta \in \mathbb{R}$) inconnu, Z une variable aléatoire ($Z \rightsquigarrow \mathcal{N}(0,1)$).

Estimation du paramètre θ

Pour estimer le paramètre θ, on fait n ($n \geq 1$) essais indépendants avec des doses de l'agent A notées $x_1, \dots, x_n$. De Y, on extrairt donc un échantillon aléatoire bernoullien : $Y_{\underline{X}_n} = (Y_{x_1}, \dots, Y_{x_n})$.

Modèle statistique associé à $\underline{Y}_n$

On se place dans un modèle statistique paramétré :

$$(\mathbb{R}^n, \otimes_{i=1}^n \mathcal{N}(\theta x_i^2, a^2)_{(\theta,a)\in\mathbb{R}^2\times]0;+\infty[})$$

La vraisemblance du modèle est :

$$L(Y_{\underline{X}_n}, a^2, \theta) = (2\pi a^2)^{-\frac{n}{2}} \times e^{\frac{\sum_{i=1}^n (Y_i - \theta x_i^2)^2}{2a^2}}$$

Régularité du modèle associé à $\underline{Y}_n$

La démonstration du livret du module 2 du cours (page 98), démontre la régularité du modèle suivant :

$$(\mathbb{R}^n, \otimes_{i=1}^n \mathcal{N}(\theta_0 + \theta_1 x_i', \sigma^2)_{(\theta_0,\theta_1,\sigma)\in\mathbb{R}^2\times]0;+\infty[})$$

En posant $\theta_0 = 0$, $\theta_1 = \theta$, $\sigma = a$ et $x_i' = x_i^2$, on retrouve le même modèle statistique associé plus haut à $\underline{Y}_n$. Ainsi le modèle de $\underline{Y}_n$ est régulier.

Calcul de l'EMV $\hat{\theta}_n$

Selon le théorème de Gauss-Markov, on a :

$$\hat{\theta}_n = \frac{\sum_{i=1}^n (x_i^2 - \overline{x_n{}^2})(Y_{x_i} - \overline{Y}_{\underline{X}_n})}{n S^2_{\underline{X}_n}}$$

avec :

$$\overline{x_n{}^2} = \frac{\sum_{i=1}^n x_i^2}{n}$$

et

$$S^2_{\underline{X}_n} = \frac{\sum_{i=1}^n (x_i^2 - \overline{x_n{}^2})^2}{n}$$

Démonstration de $\hat{\theta}_n$ en tant qu'estimateur efficace de θ

Le modèle statistique étant régulier et les distributions d'échantillonnage normales, nous savons (selon le chapitre 2 du livret du module 2 du cours) que les EMV fournissent des estimateurs consistants efficaces ou asymptotiquement efficaces.

Ainsi l'EMV $\hat{\theta}_n$ est un estimateur efficace de θ.

Calcul de la borne de Cramer-Rao pour θ

La borne Cramér-Rao exprime une borne inférieure sur la variance d'un estimateur sans biais, basée sur l'information de Fisher ($\mathcal{I}$) (selon stats.stackexchange.com).

Or on sait que ce modèle statistique est régulier, et que le livret du module 2 du cours (page 89) définit $\hat{\theta}_n$ en tant qu'estimateur efficace de θ, tel que :

$$\hat{\theta}_n \rightsquigarrow \mathcal{N}(\theta, \frac{a^2}{nS^2_{\underline{X}_n}})$$

Ainsi, on peut écrire la borne Cramér-Rao pour θ telle que :

$$\mathcal{I}_{\underline{X}_n}(\theta)^{-1} = \frac{a^2}{nS^2_{\underline{X}_n}}$$

Identification de la loi de $\hat{\theta}_n$

On a $\hat{\theta}_n$ tel que :

$$\hat{\theta}_n \rightsquigarrow \mathcal{N}(\theta_1, \frac{a^2}{nS^2_{\underline{X}_n}})$$

Calcul d'un intervalle de confiance (IC) de niveau $1-\alpha$ pour θ

Le livret du module 2 du cours (page 90), définit l'intervalle de confiance (IC) du paramètre θ tel que :

$$IC^{1-\alpha}_{Y_{\underline{X}_n}}(\theta) = \left[\hat{\theta}_n - Z_{1-\frac{\alpha}{2}}\sqrt{\frac{a^2}{nS^2_{\underline{X}_n}}}; \hat{\theta}_n + Z_{1-\frac{\alpha}{2}}\sqrt{\frac{a^2}{nS^2_{\underline{X}_n}}}\right]$$

Estimation ponctuelle du paramètre θ

On pose suppose $a = 0.5$.

Description des données

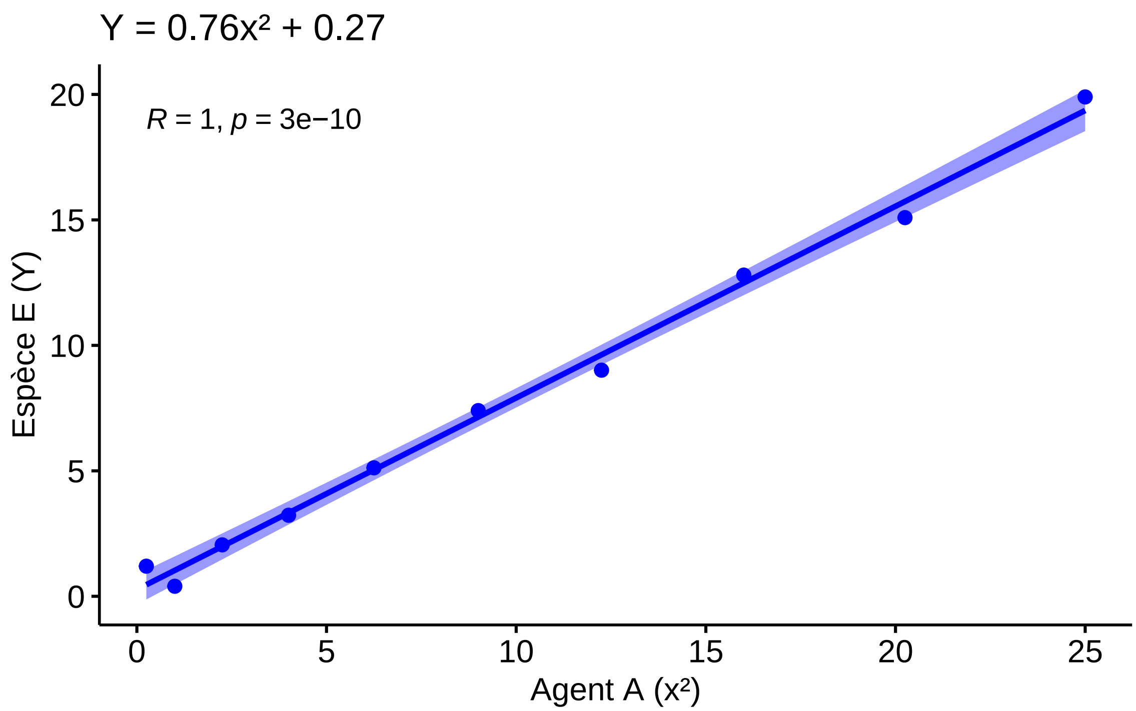
L'effectif des données est de 10.

Dans la table suivante on présente les données nécessaires à l'étude du modèle de régression linéaire :

x_i	0.50	1.0	1.50	2.00	2.50	3.0	3.50	4.0	4.50	5.0
x_i^2	0.25	1.0	2.25	4.00	6.25	9.0	12.25	16.0	20.25	25.0

y_i	1.20	0.4	2.05	3.23	5.12	7.4	9.01	12.8	15.09	19.9

Le graphique, ci-dessous illustre le modèle aléatoire suivant : $Y = \theta x^2 + aZ$, soulignant la relation linéaire entre Y et x^2, tel que $Y = ax^2 + b$ avec $a = 0.76$ et $b = 0.27$, calculés avec la méthode de corrélation de pearson et avec un intervalle de confiance de 0.95.



Avec un R si grand, on peut supposer l'intensité de liaison entre Y et x^2 très forte. En effet, on a :

	Estimate	Std. Error	t value	Pr(>|t|)
Ordonnée à l'origine	0.270	0.258	1.045	0.327
x_i^2	0.764	0.021	37.219	0.000

Calcul d'une estimation ponctuelle de θ

Il a pu être défini que :

$$\hat{\theta}_n = \frac{\sum_{i=1}^{n}(x_i^2 - \overline{x_n}^2)(Y_{x_i} - \overline{Y}_{\underline{X}_n})}{nS^2_{\underline{X}_n}}$$

Ainsi à partir de l'échantillon, comme on a :

$$nS^2_{\underline{X}_n} = \sum_{i=1}^{n}(x^2 - \overline{x_n^2}) = 656,906$$

et

$$\sum_{i=1}^{n}(x^2 - \overline{x_n^2})(Y_{x_i} - \overline{Y_{\underline{X}_n}}) = 501,653$$

On peut en déduire $\hat{\theta}_n$, tel que :

$$\hat{\theta}_n = \theta = 0,7637$$

On obtient dès lors la table suivante :

x_i	0.50	1.00	1.50	2.00	2.50	3.00	3.50	4.00	4.50	5.00
x_i^2	0.25	1.00	2.25	4.00	6.25	9.00	12.25	16.00	20.25	25.00
y_i	1.20	0.40	2.05	3.23	5.12	7.40	9.01	12.80	15.09	19.90
$\tilde{y_i} = \tilde{\theta}x^2$	0.19	0.76	1.72	3.05	4.77	6.87	9.36	12.22	15.46	19.09
$\tilde{e_i} = y_i - \tilde{y_i}$	1.01	-0.36	0.33	0.18	0.35	0.53	-0.35	0.58	-0.37	0.81

On pourra noter que l'on obtient un écart résiduel non nul, tel que :

$$\sum_{i=1}^{n} \tilde{e_i} = 2,7$$

Calcul d'un intervalle de confiance (IC) de niveau 95% pour θ

$$\left[\theta - Z_{1-\frac{\alpha}{2}} \sqrt{\frac{a^2}{nS^2_{\underline{X}_n}}} ; \theta + Z_{1-\frac{\alpha}{2}} \sqrt{\frac{a^2}{nS^2_{\underline{X}_n}}}\right] : (3)$$

Or on suppose que l'écart-type (a) est tel que $a = 0.5$, ainsi :

$$(3) \Leftrightarrow \left[0,7637 - 1,96 \times \sqrt{\frac{0,5^2}{656,906}} ; 0,7637 + 1,96 \times \sqrt{\frac{0,5^2}{656,906}}\right]$$

$$(3) \Leftrightarrow [0,7254; 0,8019]$$

Evaluation d'une évolution significative de θ

On ne souhaite ici évaluer une croissance ou une décroissance, de θ ; mais seulement savoir s'il y a une différence.

Ainsi, on souhaite savoir si l'on peut accepter au seuil de 5% l'hypothèse $H_1 : \theta = \theta_{pass} \Leftrightarrow H_0 : \theta = 0,9$ contre $H_1 : \theta \neq 0,9$.

Test d'hypothèses

$$W^{\alpha}_{Y_{\underline{X}_n}}(\theta) = \left\{ \mid \hat{\theta}_n - 0,9 \mid > Z_{1-\frac{\alpha}{2}} \sqrt{\frac{a^2}{nS^2_{\underline{X}_n}}} \right\} : (4)$$

$$(4) \Leftrightarrow W^{\alpha}_{Y_{\underline{X}_n}}(\theta) = \left\{ \mid \hat{\theta}_n - 0,9 \mid > 1,96 \times \sqrt{\frac{0,5^2}{656,906}} \right\}$$

$$(4) \Leftrightarrow W^{\alpha}_{Y_{\underline{X}_n}}(\theta) = \left\{ \mid \hat{\theta}_n - 0,9 \mid > 0,038 \right\}$$

Observations On observe :

$$| \hat{\theta}_n - 0,9 | = | 0,7637 - 0,9 | = | -0.1363 | = 0.1363 : (5)$$
$$(5) \Leftrightarrow | \hat{\theta}_n - 0,9 | \in W^{\alpha}_{Y_{\underline{X}_n}}$$

Décision Au seuil de de 5%, on rejete H_0 jusqu'à preuve du contraire.

Conlcusion On en conclue que la valeur du paramètre θ a évolu significativement par rapport aux mesures précédentes.