

AES MISASHS L3
Statistique-Inférence

C. Joutard

Université Paul Valéry

Année universitaire 2014-2015

Organisation

1) 3h Cours-TD par semaine

2) Informations et documents sur le site

<http://www.univ-montp3.fr/miap/ens/>

cliquer sur le lien “MisashsAES” dans la colonne de gauche puis sur “Statistique - Inférence” (qui se trouve dans la rubrique “AES Misashs L3”).

Chapitre I. Loix limites de la Statistique et Estimation

1) Loi des grands nombres et estimation ponctuelle

1.1. Loi des grands nombres

Soient X_1, X_2, \dots, X_n , n variables aléatoires indépendantes associées aux répétitions d'une même expérience aléatoire X telle que $\mathbb{E}(X) = \mu$ alors :

La moyenne des observations (aussi appelée “moyenne empirique”) est aussi proche que possible de la vraie valeur μ à condition que n soit grand.

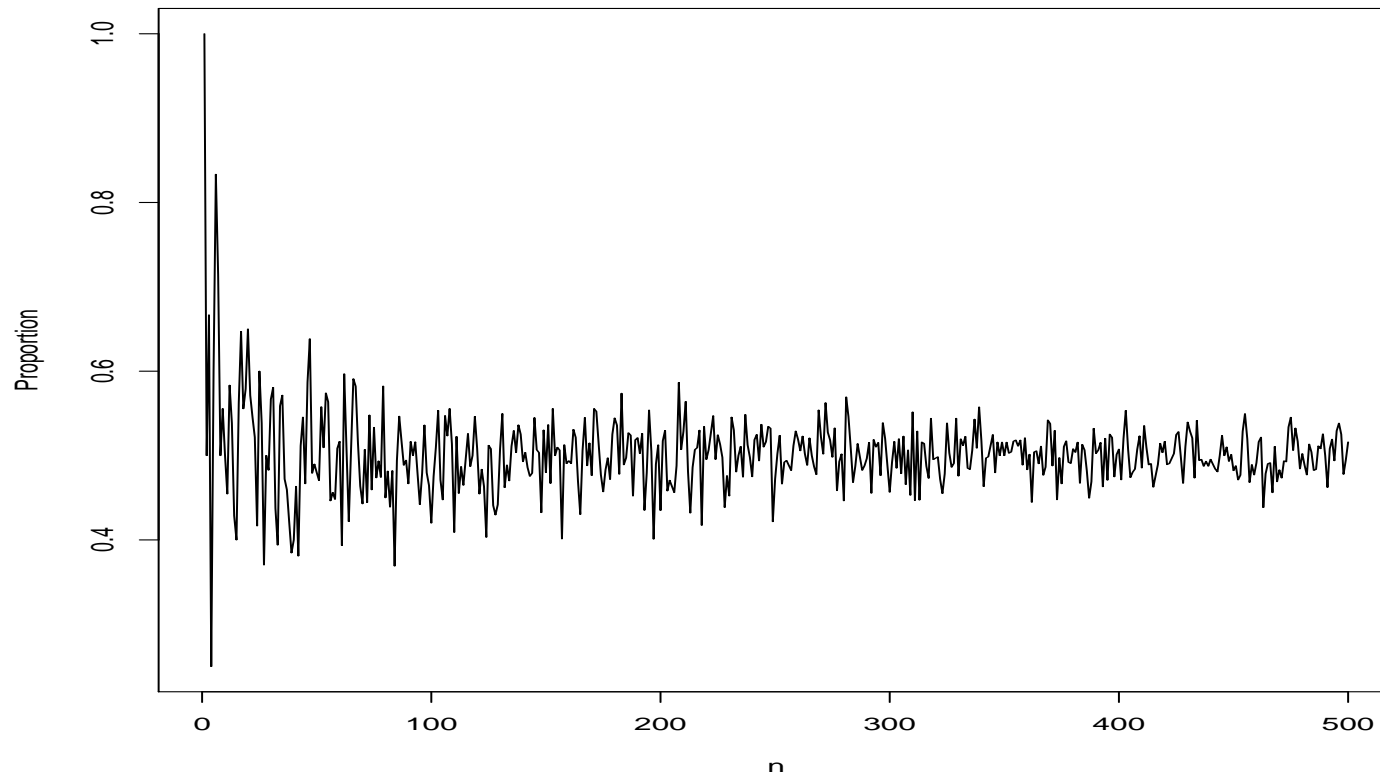
$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i \xrightarrow{n \text{ grand}} \mu$$

Exemple 1 : Lancer d'une pièce de monnaie. On lance une pièce équilibrée. Le résultat "*pile/face*" est codé par "0/1". La pièce étant équilibrée, les probabilités sont: $P(\text{"Pile"}) = P(X = 0) = 1/2$ et $P(\text{"Face"}) = P(X = 1) = 1/2$ où la variable aléatoire X est le résultat du lancer. L'espérance de X est donc égale à $\mu = 1/2$.

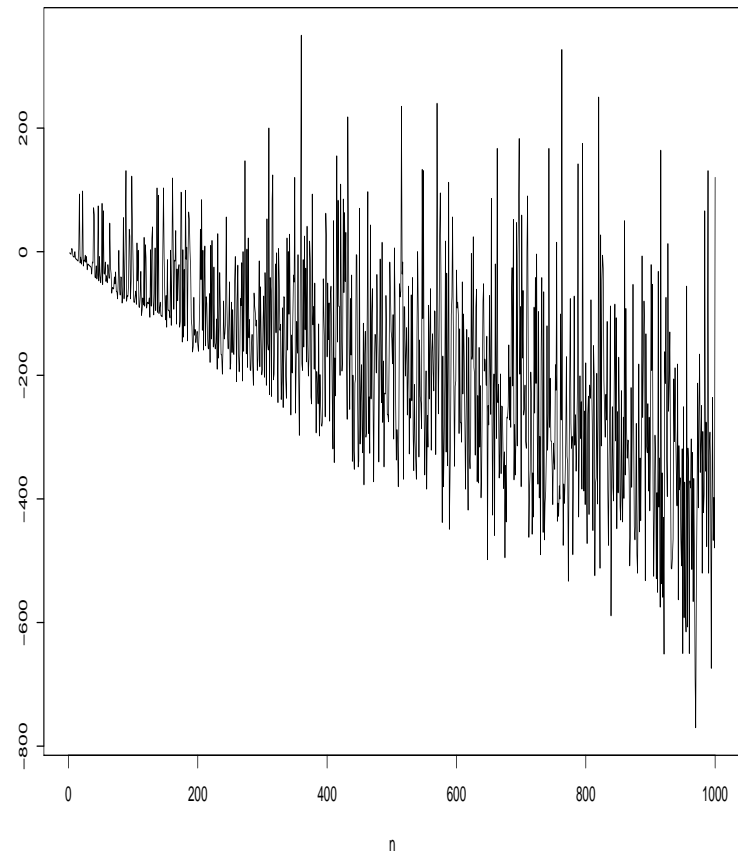
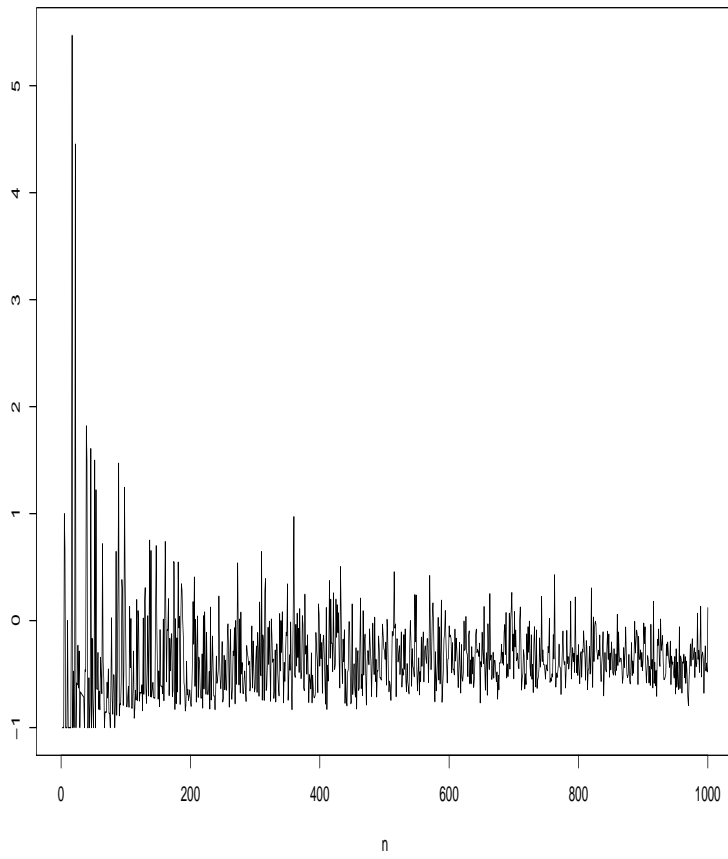
D'après la loi des grands nombres, si on lance la pièce un grand nombre de fois (on note n le nombre de lancers) et qu'on relève à chaque lancer le résultat, on obtiendra alors une proportion de "faces" proche de $\mu = 1/2$.

Exemple 2 : Jeu de hasard : jeu consistant à miser 1 euro et pouvant rapporter soit 10 euros soit 100 euros. Les chances d'obtenir ces deux gains sont respectivement de $1/30$ et $1/300$. Soit la variable X correspondant au gain d'un joueur après un essai. Les valeurs possibles de X sont donc $\{-1, 9, 99\}$. Un joueur a une espérance de gain égale à $\mu = -1 * \frac{289}{300} + 9 * \frac{1}{30} + 99 * \frac{1}{300} = -\frac{1}{3}$. Par conséquent, si un joueur joue un très grand nombre de fois, il est (pratiquement) sûr de perdre de l'argent.

Exemple 1 :



Exemple 2 :



Propriétés de la moyenne $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$

où X_1, X_2, \dots, X_n sont les répétitions indépendantes d'une même expérience aléatoire X telle que $\mathbb{E}(X) = \mu$ et $V(X) = \sigma^2$:

- Son espérance est : $\mathbb{E}(\bar{X}) = \mu$.
- Sa variance est : $V(\bar{X}) = \frac{\sigma^2}{n}$.
- Son écart-type est : $\sigma(\bar{X}) = \frac{\sigma}{\sqrt{n}}$.

La dispersion de la moyenne se réduit quand n grandit : c'est la loi des grands nombres.

La loi de bernoulli et le sondage

- Soit X une expérience aléatoire à 2 états (codés 1/0) :

$$P(X = 1) = p; P(X = 0) = 1 - p$$

On sait que $X \sim Ber(p)$ avec :

- Son espérance mathématique $\mathbb{E}(X) = p$.
 - Sa variance $V(X) = \mathbb{E}(X - \mathbb{E}(X))^2 = p(1 - p)$.
 - Et son écart-type $\sigma(X) = \sqrt{p(1 - p)}$.
- La même expérience (de Bernoulli) répétée n fois de façon indépendante X_1, X_2, \dots, X_n alors :

L'espérance mathématique de $\sum_{i=1}^n X_i$ est np .

L'espérance de la somme est la somme des espérances.

La moyenne de loi de Bernoulli :

- Son espérance est :

$$\mathbb{E}(\bar{X}) = p.$$

- Sa variance est :

$$V(\bar{X}) = \frac{p(1-p)}{n}.$$

- Son écart-type est :

$$\sigma(\bar{X}) = \sqrt{\frac{p(1-p)}{n}}.$$

La dispersion de la moyenne (ici la proportion de 1) se réduit quand n grandit : c'est la **loi des grands nombres**.

La loi des grands nombres pour la loi de Bernoulli

Soient X_1, X_2, \dots, X_n , n répétitions indépendantes d'une même expérience aléatoire de Bernoulli ($P(X = 1) = p$) alors :

La proportion de 1 est aussi proche que possible de p
à condition que n soit grand

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \xrightarrow{n \text{ grand}} p$$

1.2. Estimation ponctuelle

D'après ses propriétés, la moyenne $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ † est donc un bon candidat pour approcher le paramètre inconnu μ .

- La variable aléatoire \bar{X} sera appelée l'**Estimateur** du paramètre μ (notation: $\hat{\mu} = \bar{X}$).
- Au vu d'observations, la valeur prise par \bar{X} et notée \bar{x} sera appelée une **estimation** du paramètre μ .
- En pratique il y a UN (voire deux ou trois) Estimateur naturel du paramètre inconnu, mais il y a toujours une infinité d'estimations possibles de ce paramètre.

† X_1, X_2, \dots, X_n sont n répétitions indépendantes d'une même expérience aléatoire X telle que $\mathbb{E}(X) = \mu$ et $V(X) = \sigma^2$.

Forme générale d'un estimateur et propriétés

Soient n répétitions (indépendantes) X_1, X_2, \dots, X_n d'une expérience aléatoire et T_n une combinaison (ou fonction) de ces répétitions.

T_n sera un bon candidat pour approcher un paramètre inconnu θ et donc un **Estimateur** raisonnable de θ (on pourra donc écrire $T_n = \hat{\theta}$) si :

- L'espérance mathématique de l'estimateur est aussi proche que possible du paramètre inconnu θ , idéalement on souhaite que

$$\mathbb{E}(T_n) = \theta$$

et on dira que T_n est un estimateur **sans biais** de θ ; mais on peut se contenter de $\mathbb{E}(T_n) \xrightarrow{n \text{ grand}} \theta$.

- La variance de l'estimateur diminue avec le nombre de répétitions :

$$V(T_n) \xrightarrow{n \text{ grand}} 0$$

Exemple : Cas de n répétitions indépendantes X_1, X_2, \dots, X_n d'une même expérience aléatoire X telle que $\mathbb{E}(X) = \mu$ et $V(X) = \sigma^2$.

- Si μ est inconnu alors la moyenne $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ est un estimateur **sans biais** de μ , noté $\hat{\mu}$.
- Si μ est connu et σ^2 inconnu alors la moyenne des dispersions $\frac{1}{n} \sum_{i=1}^n (X_i - \mu)^2$ est un estimateur **sans biais** de σ^2 , noté $\hat{\sigma}_\mu^2$.
- Si μ est inconnu et σ^2 inconnu alors la **variance empirique** $\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$ est un estimateur **biaisé** de σ^2 , noté $\hat{\sigma}^2$.
- Si μ est inconnu et σ^2 inconnu alors $\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$ est un estimateur **sans biais** de σ^2 que l'on notera $\hat{\sigma}_{SB}^2$.

Exercice : On suppose avoir observé n variables aléatoires Y_1, Y_2, \dots, Y_n et on propose 3 estimateurs d'un paramètre θ : T_1, T_2 et T_3 ayant les propriétés suivantes (λ est un autre paramètre inconnu) :

$$\begin{aligned} E(T_1) &= \theta + \lambda & \text{et} & & V(T_1) &= (\theta * \lambda)/n \\ E(T_2) &= \theta + \lambda/n & \text{et} & & V(T_2) &= (\theta * \lambda)/n \\ E(T_3) &= \theta & \text{et} & & V(T_3) &= \lambda \end{aligned}$$

Donner les propriétés :

- de biais ("estimateur sans biais" ; "son biais diminue quand n , le nombre d'observations grandit")
- et de variance ("sa variance diminue quand n , le nombre d'observations grandit")

Lequel des 3 est-il raisonnable de garder?

Estimation ponctuelle dans le cas de Bernoulli

Soient X_1, X_2, \dots, X_n , n répétitions indépendantes d'une même expérience aléatoire de Bernoulli ($P(X = 1) = p$, p paramètre inconnu) alors :

- La variable aléatoire \bar{X} (ici la proportion ou la fréquence de 1) sera donc un **Estimateur sans biais** du paramètre p ; $\hat{p} = \bar{X}$.
- Après avoir effectué le sondage, la valeur prise par \bar{X} et notée \bar{x} sera donc une **estimation** du paramètre p .
- En pratique on a toujours le même Estimateur du paramètre inconnu p , mais chaque sondage pratiqué amène une estimation différente de ce paramètre.

2) Théorème central limite et estimation par intervalle

2.1. Définition d'une variable centrée, réduite

- Variable centrée : $X - \mathbb{E}(X)$
- Variable centrée, réduite : $\frac{X - \mathbb{E}(X)}{\sigma(X)}$

Une variable centrée réduite a pour espérance 0 et écart-type 1.

Pour la moyenne de n répétitions indépendantes d'une même expérience aléatoire d'espérance μ et de variance σ^2 :

$$\frac{\bar{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}}$$
 est une variable centrée réduite.

Pour la moyenne de répétitions de même loi de Bernoulli indépendantes de paramètre p :

$$\frac{\bar{X} - p}{\sqrt{\frac{p(1-p)}{n}}}$$
 est une variable centrée réduite.

2.2. Théorème central limite

Soient X_1, X_2, \dots, X_n , n répétitions indépendantes d'une même expérience aléatoire alors :

La moyenne se comporte comme une loi normale

à condition que n soit grand

donc la moyenne centrée et réduite se comporte comme une $\mathcal{N}(0, 1)$.

$$\frac{\bar{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \xrightarrow{n \text{ grand}} \mathcal{N}(0, 1)$$

Le sondage

Soient X_1, X_2, \dots, X_n , n répétitions indépendantes d'une même expérience aléatoire de Bernoulli ($P(X = 1) = p$) alors :

La proportion de 1 se comporte comme une loi normale

à condition que n soit grand

donc la proportion de 1 centrée et réduite se comporte comme une $\mathcal{N}(0, 1)$.

$$\frac{\bar{X} - p}{\sqrt{\frac{p(1-p)}{n}}} \xrightarrow{n \text{ grand}} \mathcal{N}(0, 1)$$

2.3. Intervalle de dispersion

Terminologie de base : on cherche à construire un intervalle de grande probabilité de la réalisation d'une expérience aléatoire dont la loi de probabilité est connue.

Pour cela on se fixe une faible probabilité α , appelée "niveau" (en pratique α vaut 1%, 5% parfois 10%, 0.1%) et on construit un

Intervalle de dispersion à $1 - \alpha$

(l'expérience aléatoire a une probabilité $1 - \alpha$ de se réaliser dans l'intervalle).

Exemple 1 : si on lance 500 fois une même pièce, on cherche l'intervalle de dispersion à 95% du nombre de faces.

$$P(\text{Nbrefaces} \in [B_1; B_2]) = 95\%$$

Est ce $[220, 280]$? ou alors $[240, 260]$? ou un autre intervalle?

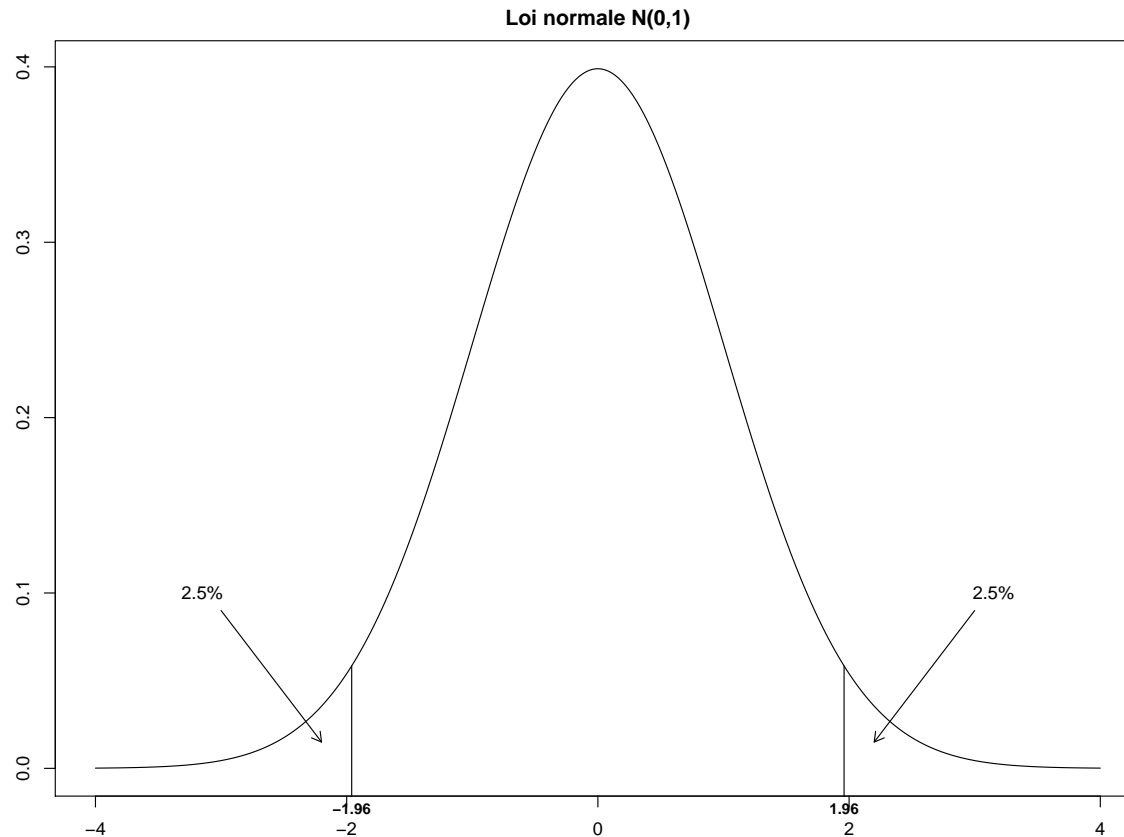
Exemple 2 : Dans un échantillon de 200 étudiants, on s'intéresse à l'intervalle de dispersion à 99% du nombre de fumeurs.

Exemple 3 : Une chaîne de supermarché décide de supprimer dans ses prix toutes références aux centimes d'euros par l'arrondi suivant :
 $\{0; 1; 2\}$ donne 0 ; $\{3; 4\}$ donne 5.

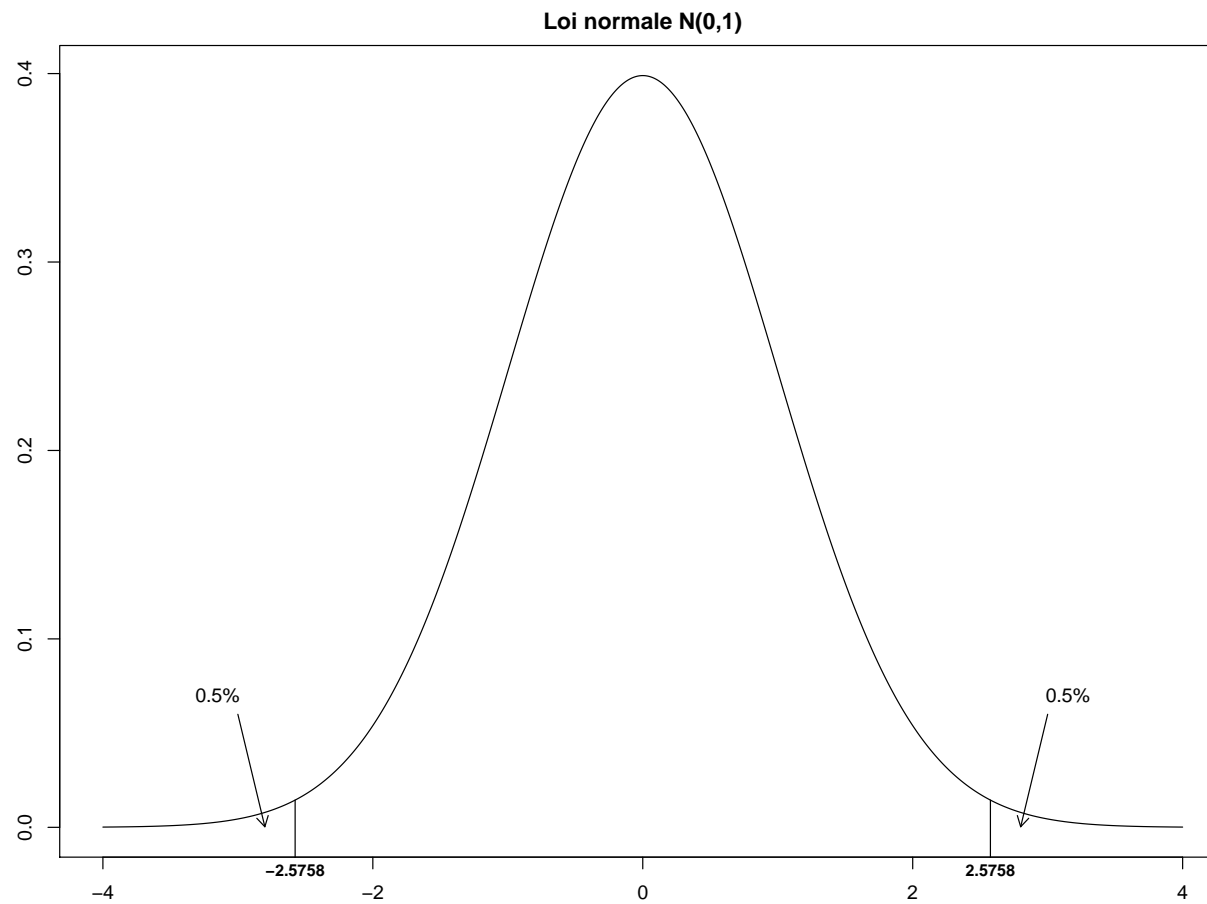
A quoi peut-elle s'attendre après la vente de 20000 produits?

Intervalle de dispersion de la loi normale

Soit Z une variable aléatoire de loi normale $\mathcal{N}(0, 1)$ alors :



$P(-1.96 \leq Z \leq 1.96) = 0.95$. Donc $[-1.96; 1.96]$ est un intervalle de dispersion (ID) à 95%.



$P(-2.5758 \leq Z \leq 2.5758) = 0.99$. Donc $[-2.5758; 2.5758]$ est un intervalle de dispersion (ID) à 99%.

De façon générale, on notera

$$ID_{1-\alpha}(Z) = [-l_{\frac{\alpha}{2}}; l_{\frac{\alpha}{2}}]$$

l'**intervalle de dispersion** à $1 - \alpha$ d'une v.a. Z de loi normale $\mathcal{N}(0, 1)$

$$P(-l_{\frac{\alpha}{2}} \leq Z \leq l_{\frac{\alpha}{2}}) = 1 - \alpha$$

Intervalle de dispersion d'une somme de Bernoulli, notée S_n

Soient X_1, X_2, \dots, X_n , n réalisations indépendantes d'une même expérience aléatoire de Bernoulli ($P(X = 1) = p_0$, p_0 connu) alors on sait grâce au TCL que :

$$\frac{\sum_{i=1}^n X_i - np_0}{\sqrt{np_0(1-p_0)}} = \frac{S_n - np_0}{\sqrt{np_0(1-p_0)}} \xrightarrow{n \text{ grand}} \mathcal{N}(0, 1)$$

Et on sait également que si Z est une variable aléatoire de loi $\mathcal{N}(0, 1)$ alors :

$$P(-l_{\frac{\alpha}{2}} \leq Z \leq l_{\frac{\alpha}{2}}) = 1 - \alpha$$

Donc

$$P\left(-l_{\frac{\alpha}{2}} \leq \frac{S_n - np_0}{\sqrt{np_0(1-p_0)}} \leq l_{\frac{\alpha}{2}}\right) \approx 1 - \alpha$$

Et,

$$P\left(np_0 - l_{\frac{\alpha}{2}} \sqrt{np_0(1-p_0)} \leq S_n \leq np_0 + l_{\frac{\alpha}{2}} \sqrt{np_0(1-p_0)}\right) \approx 1 - \alpha$$

$$ID_{1-\alpha}(S_n) = \left[np_0 - l_{\frac{\alpha}{2}} \sqrt{np_0(1-p_0)}, np_0 + l_{\frac{\alpha}{2}} \sqrt{np_0(1-p_0)} \right]$$

est donc un **ID** de S_n de probabilité approximative $(1 - \alpha)$.

Exemple : Sur 500 naissances, le nombre de garçons sera compris à 95% entre :

$$\begin{aligned} & \left[500 \times \frac{1}{2} - 1.96 \sqrt{500 \times \frac{1}{2} \times \left(1 - \frac{1}{2}\right)}, 500 \times \frac{1}{2} + 1.96 \sqrt{500 \times \frac{1}{2} \times \left(1 - \frac{1}{2}\right)} \right] \\ & = [250 - 21.91, 250 + 21.91] \approx [228, 272] \end{aligned}$$

L'intervalle à 99% serait de [221, 279].

Intervalle de dispersion d'une moyenne puis d'une somme de v.a.

Soient X_1, X_2, \dots, X_n , n réalisations indépendantes d'une même expérience aléatoire d'espérance μ_0 et de variance σ_0^2 connues alors (TCL) :

$$\frac{\bar{X} - \mu_0}{\sqrt{\frac{\sigma_0^2}{n}}} \xrightarrow{n \text{ grand}} \mathcal{N}(0, 1)$$

Donc,

$$P\left(-l_{\frac{\alpha}{2}} \leq \frac{\bar{X} - \mu_0}{\sqrt{\frac{\sigma_0^2}{n}}} \leq l_{\frac{\alpha}{2}}\right) \approx 1 - \alpha$$

puis

$$P\left(\mu_0 - l_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_0^2}{n}} \leq \bar{X} \leq \mu_0 + l_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_0^2}{n}}\right) \approx 1 - \alpha$$

$$\text{D'où } ID_{1-\alpha}(\bar{X}) = \left[\mu_0 - l_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_0^2}{n}}; \mu_0 + l_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_0^2}{n}} \right].$$

Pour la somme $S_n = \sum_{i=1}^n X_i$, on obtient par le même raisonnement,

$$ID_{1-\alpha}(S_n) = \left[n\mu_0 - l_{\frac{\alpha}{2}} \sqrt{n\sigma_0^2}; n\mu_0 + l_{\frac{\alpha}{2}} \sqrt{n\sigma_0^2} \right].$$

2.4. Estimation par intervalle : l'intervalle de confiance

Notion élémentaire d'intervalle de confiance

Terminologie de base : on cherche à construire un intervalle de grande probabilité d'une grandeur particulière inconnue (en général le paramètre d'intérêt d'une loi de probabilité).

Pour cela on se fixe une faible probabilité α , appelée "risque" (en pratique α vaut 1%, 5% parfois 10%, 0.1%) et on construit un

Intervalle de confiance à $1 - \alpha$

que l'on notera par la suite $IC_{1-\alpha}$ (paramètre), et qui représente un intervalle (dont les bornes sont aléatoires) qui a une probabilité $1 - \alpha$ de contenir le paramètre inconnu recherché.

Exemple 1 : à la veille d'une élection, on réalise un sondage pour connaître le plus précisément possible le paramètre associé au score d'un candidat.

$$P(\text{score} \in [B_1, B_2]) = 95\%.$$

Exemple 2 : on mesure les performances de n individus ($n = 200$) afin de proposer un intervalle pour l'ensemble de la population.

$$P(\text{performance} \in [B_1, B_2]) = 99\%.$$

Les bornes B_1, B_2 de l'intervalle sont des v.a. et seront calculées en pratique à l'aide des observations.

Intervalle de confiance du paramètre p d'une loi de Bernoulli

Soient X_1, X_2, \dots, X_n , n répétitions indépendantes d'une même expérience aléatoire X de Bernoulli ($X \sim \text{Ber}(p)$) alors on sait que :

$$\frac{\bar{X} - p}{\sqrt{\frac{p(1-p)}{n}}} \xrightarrow{n \text{ grand}} \mathcal{N}(0, 1)$$

Et on sait également que si Z est une variable aléatoire de loi $\mathcal{N}(0, 1)$ alors :

$$P(-l_{\frac{\alpha}{2}} \leq Z \leq l_{\frac{\alpha}{2}}) = 1 - \alpha.$$

Donc

$$P\left(-l_{\frac{\alpha}{2}} \leq \frac{\bar{X} - p}{\sqrt{\frac{p(1-p)}{n}}} \leq l_{\frac{\alpha}{2}}\right) \approx 1 - \alpha.$$

Et,

$$P \left(\bar{X} - l_{\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}} \leq p \leq \bar{X} + l_{\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}} \right) \approx 1 - \alpha$$

$$P \left(\bar{X} - l_{\frac{\alpha}{2}} \sqrt{\frac{\bar{X}(1-\bar{X})}{n}} \leq p \leq \bar{X} + l_{\frac{\alpha}{2}} \sqrt{\frac{\bar{X}(1-\bar{X})}{n}} \right) \approx 1 - \alpha$$

car p inconnu est remplacé par son estimateur $\hat{p} = \bar{X}$.

D'où

$$IC_{1-\alpha}(p) = \left[\bar{X} - l_{\frac{\alpha}{2}} \sqrt{\frac{\bar{X}(1-\bar{X})}{n}}, \bar{X} + l_{\frac{\alpha}{2}} \sqrt{\frac{\bar{X}(1-\bar{X})}{n}} \right]$$

est donc un intervalle de confiance de p de probabilité approximative $(1 - \alpha)$.

Exemple :

$$n = 1000$$

\bar{x}	$\sqrt{\frac{\bar{x}(1-\bar{x})}{n}}$	IC à 95%	IC à 99%
0.2	0.01265	[0.1752 , 0,2248]	[0.1674 , 0.2326]
0.5	0.01581	[0.4690 , 0.5310]	[0.4593 , 0.5407]
0.9	0.00949	[0.8814 , 0.9186]	[0.8756 , 0.9244]

$$n = 500$$

\bar{x}	$\sqrt{\frac{\bar{x}(1-\bar{x})}{n}}$	IC à 95%	IC à 99%
0.3	0.0205		
0.4			

Intervalle de confiance du paramètre μ d'une loi gaussienne $\mathcal{N}(\mu, \sigma^2)$

Avec X_1, X_2, \dots, X_n , n répétitions indépendantes d'une même expérience aléatoire de loi gaussienne $\mathcal{N}(\mu, \sigma^2)$, on sait que :

$$\frac{\bar{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \sim \mathcal{N}(0, 1)$$

Et on sait également que si Z est une variable aléatoire $\mathcal{N}(0, 1)$ alors :

$$P(-l_{\frac{\alpha}{2}} \leq Z \leq l_{\frac{\alpha}{2}}) = 1 - \alpha.$$

Donc

$$P\left(-l_{\frac{\alpha}{2}} \leq \frac{\bar{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \leq l_{\frac{\alpha}{2}}\right) = 1 - \alpha.$$

Et,

$$P \left(\bar{X} - l_{\frac{\alpha}{2}} \sqrt{\frac{\sigma^2}{n}} \leq \mu \leq \bar{X} + l_{\frac{\alpha}{2}} \sqrt{\frac{\sigma^2}{n}} \right) = 1 - \alpha$$

D'où

$$IC_{1-\alpha}(\mu) = \left[\bar{X} - l_{\frac{\alpha}{2}} \sqrt{\frac{\sigma^2}{n}}, \bar{X} + l_{\frac{\alpha}{2}} \sqrt{\frac{\sigma^2}{n}} \right]$$

est donc un intervalle de confiance de μ de probabilité $(1 - \alpha)$ lorsque σ^2 est connu.

Dans le cas où σ est inconnu, on remplacera σ^2 par un estimateur : la variance empirique des observations $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$.

Ce qui nécessite un grand nombre d'observations n .

Chapitre II. Tests statistiques

1) Introduction et vocabulaire

1.1. Notion élémentaire de test statistique

Terminologie de base : on cherche à construire un test d'une hypothèse H_0 (appelée hypothèse nulle) contre une hypothèse H_1 (appelée hypothèse alternative) dans le but de rejeter l'hypothèse H_0 .

Construire le test de H_0 contre H_1

C'est adopter une règle de décision qui amène :

à rejeter H_0 ou à ne pas rejeter H_0 .

Exemple 1 : avant son entrée en campagne, la popularité d'un candidat était stabilisée à 40%. Un récent sondage lui donne une popularité de 43%. Est-ce que la popularité du candidat a effectivement augmenté ou est-ce dû à une fluctuation du sondage?

On construira le test de

- l'hypothèse H_0 “*status quo*” contre
- l'hypothèse alternative H_1 “*augmentation de la popularité*”.

Exemple 2 : un fabricant de composant assure la qualité de son produit par la phrase suivante “*la fiabilité de ma production est supérieure ou égale à 99%*”. Lors d’un contrôle, on relève 1.09% de pièces défectueuses sur un échantillon de 1000 pièces. Doit-on supprimer l’accréditation au fabricant?

On construira le test de :

- l’hypothèse H_0 : “*le taux de pièces défectueuses est de 1%*” contre
- l’hypothèse alternative H_1 : “*le taux de pièces défectueuses est $> 1%$* ”.

Exemple 3 : on connaît les résultats d’un test (éducatif, psychologique, ...) sur l’ensemble de la population. On fait alors pratiquer ce test sur un groupe particulier et on observe sur ce groupe une augmentation des résultats. Est-ce que ce groupe a effectivement de meilleurs résultats ou est-ce dû à une fluctuation d’échantillonnage?

On construira le test de :

- l’hypothèse H_0 : “*le résultat au test est identique à celui de la population générale*”.
- l’hypothèse alternative H_1 : “*le résultat au test est supérieur*”.

1.2. Erreurs du test statistique

- On peut se tromper en déclarant H_1 vraie alors que H_0 est vraie : c'est l'erreur de 1^{re} espèce.
- On peut se tromper en déclarant H_0 vraie alors que H_1 est vraie : c'est l'erreur de 2^e espèce.

	Choix de H_0	Choix de H_1
H_0 vraie	Décision juste	Erreur de 1 ^{re} espèce
H_1 vraie	Erreur de 2 ^e espèce	Décision juste

En pratique, l'erreur de 1^{re} espèce ou niveau, notée α est fixée par l'utilisateur et on construit donc

un test de H_0 contre H_1 de niveau α .

Dans les exemples précédents l'erreur de 1^{re} espèce correspond à :

Exemple 1 : P(croire à une “augmentation de la popularité” alors qu’il n’en est rien).

Exemple 2 : P(supprimer l'accréditation au fabricant alors que le taux de pièces défectueuses est conforme).

Exemple 3 : P(penser à une amélioration due au groupe alors qu’il n’en est rien).

2) Test dans le cas gaussien (variable continue)

2.1. Test du paramètre d'espérance d'une loi gaussienne $\mathcal{N}(\mu, \sigma^2)$ de l'hypothèse $H_0 : \mu = \mu_0$ contre $H_1 : \mu > \mu_0$

Exemple 3 : on connaît les résultats d'un test (éducatif, psychologique, ...) sur l'ensemble de la population. On fait alors pratiquer ce test sur un groupe particulier et on observe sur ce groupe une augmentation des résultats. Est-ce que ce groupe a effectivement de meilleurs résultats ou est ce dû à une fluctuation du sondage?

On construira le test de :

- l'hypothèse H_0 : *“le résultat au test est identique à celui de la population générale”*
- l'hypothèse alternative H_1 : *“le résultat au test est supérieur”*

Exemple 3 : Les résultats d'un test (éducatif, psychologique, ...) sur l'ensemble de la population sont supposés suivre une loi $\mathcal{N}(\mu_0 = 100; \sigma^2)$. On fait alors pratiquer ce test sur un groupe particulier de 80 individus et on observe sur ce groupe une augmentation moyenne des résultats : sur ces 80 individus, la moyenne au test est de 110. Les résultats du groupe sont toujours supposés gaussiens $\mathcal{N}(\mu_{groupe} = 100; \sigma^2)$, avec la même variance σ^2 .
Y-a-t-il une augmentation significative entre les résultats du groupe et la population totale?

On construira le test de :

- l'hypothèse $H_0 : \mu_{groupe} = 100$.
- l'hypothèse alternative $H_1 : \mu_{groupe} > 100$.

Test de l'hypothèse $H_0 : “\mu = \mu_0”$ **contre** $H_1 : “\mu > \mu_0”$

Si les n observations sont gaussiennes $\mathcal{N}(\mu, \sigma^2)$ alors leur moyenne \bar{X} est gaussienne $\mathcal{N}(\mu, \frac{\sigma^2}{n})$,

et on rejettera l'hypothèse H_0 si \bar{X} est grand par rapport à μ_0 .

Se plaçant sous H_0 vraie (“ $\mu = \mu_0$ ”), on a :

$$\frac{\bar{X} - \mu_0}{\sqrt{\frac{\sigma^2}{n}}} \sim \mathcal{N}(0, 1).$$

Donc,

$$P \left(\bar{X} \leq \mu_0 + l_\alpha \sqrt{\frac{\sigma^2}{n}} \right) = 1 - \alpha$$

D'où le test de $H_0: \mu = \mu_0$ contre $H_1: \mu > \mu_0$, d'erreur de première espèce (ou niveau) α : Rejet de H_0 si

$$\bar{x} > \mu_0 + l_\alpha \sqrt{\frac{\sigma^2}{n}}$$

C'est un test unilatéral.

Mais le fait de connaître la moyenne \bar{x} ne permet aucune conclusion, la connaissance de la valeur de σ est fondamentale.

Exemple 3 : $\mu_0 = 100$; $n = 80$; pour $\alpha = 5\%$; $l_\alpha = 1.6449$

Rejet de H_0 si $\bar{X} > 100 + 0.184 \times \sigma$.

Donc si $\sigma = 40$, on rejettera si $\bar{x} > 107.36$ et si $\sigma = 80$, on rejettera si $\bar{x} > 114.72$. Dans l'énoncé, il est donné que $\bar{x}_{\text{groupe}} = 110$, on ne peut donc pas encore conclure puisque le paramètre σ n'est pas donné!

- Cas 1 : σ connu. Il n'y a donc pas de difficulté, mais ce n'est pas un cas réaliste en pratique.
- Cas 2 : σ inconnu. On remplacera σ^2 par la variance empirique des observations $\widehat{\sigma^2} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$.

En contre-partie, il est nécessaire que le nombre d'observations n soit grand.

En pratique, il est donc nécessaire d'avoir conservé les données et il est conseillé de prendre une erreur de 1^{re} espèce plus faible que lorsque l'on connaît σ .

- Cas 3 (fréquemment rencontré dans la littérature) : σ inconnu . On remplace σ^2 par $\widehat{\sigma^2} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$ qui est un estimation (sans biais) du paramètre σ^2 .

Sous l'hypothèse H_0 la statistique de test devient une loi de Student à $(n - 1)$ degrés de liberté.

$$\frac{\bar{X} - \mu_0}{\sqrt{\frac{\widehat{\sigma^2}}{n}}} \sim S_{n-1}$$

En pratique, il n'est pas nécessaire que le nombre d'observations n soit grand mais cela nécessite une grande confiance en l'hypothèse de normalité des observations.

**2.2. Test du paramètre d'espérance d'une loi gaussienne $\mathcal{N}(\mu, \sigma^2)$
de l'hypothèse $H_0 : \mu = \mu_0$ contre $H_1 : \mu \neq \mu_0$**

Si H_0 est vraie alors :

$$\frac{\bar{X} - \mu_0}{\sqrt{\frac{\sigma^2}{n}}} \sim \mathcal{N}(0, 1).$$

Donc

$$P \left(\mu_0 - l_{\frac{\alpha}{2}} \sqrt{\frac{\sigma^2}{n}} \leq \bar{X} \leq \mu_0 + l_{\frac{\alpha}{2}} \sqrt{\frac{\sigma^2}{n}} \right) = 1 - \alpha$$

D'où le test de $H_0: \mu = \mu_0$ contre $H_1: \mu \neq \mu_0$, d'erreur de première espèce (ou niveau) α :

Rejet de H_0 si

$$\bar{x} < \mu_0 - l_{\frac{\alpha}{2}} \sqrt{\frac{\sigma^2}{n}} \quad \text{ou} \quad \bar{x} > \mu_0 + l_{\frac{\alpha}{2}} \sqrt{\frac{\sigma^2}{n}}$$

C'est un test bilatéral. Si σ est inconnu, on remplace σ^2 par la variance empirique des observations $\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$.

Exemple: Dans un magasin d'habillement, on a noté ces dernières années le prix moyen d'un article du rayon "enfant" qui est de 18 euros. La direction souhaite communiquer sur sa politique de promotion dans ce rayon par la baisse du prix moyen d'un article. Pour cela, elle relève le prix de 121 articles du rayon (pris au hasard).

1. Construire le test statistique associé à cette démarche.

Le prix d'un article du rayon "enfant" est une variable aléatoire continue. Le test statistique associé à cette démarche est le test sur l'espérance de cette variable :

$H_0: \mu = 18$ contre $H_1: \mu < 18$.

La variable \bar{P} suit une loi normale et on a sous H_0 : $\frac{\bar{P}-18}{\sqrt{\frac{\sigma^2}{n}}}$ suit $\mathcal{N}(0, 1)$ et

on rejettera H_0 si $\bar{P} < 18 - l_\alpha \sigma / \sqrt{n}$,

avec $l_\alpha = 2.3263$.

2. Calculer la zone de rejet de ce test, étant donné l'information suivante : sur les 121 articles on a calculé que la dispersion $\sum_{i=1}^{121} (p_i - \bar{p})^2 = 5929$ où p_i est le prix de l'article i et \bar{p} le prix moyen observé de ces 121 articles.

Comme on ne connaît pas σ^2 , on peut l'estimer par "la dispersion / n" soit $5929/121 = 49$ donc σ est estimé par 7. La borne de rejet est :

$$18 - l_{\alpha}\sigma/\sqrt{n} = 18 - 2.3263 * 7/11 = 16.51963$$

et la zone de rejet est donc : Rejet de H_0 si $\bar{P} < 16.51963$.

3. Sur ces 121 articles le prix moyen observé au mois de mai 2008 est de 16.8 euros. Quel est la décision à prendre concernant ce test statistique? Que proposeriez-vous alors à la direction?

Au mois de mai, le prix moyen observé est de 16.8 euros; on ne rejette donc pas H_0 (le prix moyen ne baisse pas de façon significative) et on proposera alors à la direction d'accentuer sa politique de promotion.

4. Sur ces 121 articles le prix moyen observé au début du mois de juin 2008 est de 16.2 euros. Quel est la décision à prendre concernant ce test statistique?

Au mois de juin, le prix moyen observé est de 16.2 euros; on rejette donc H_0 (le prix moyen baisse de façon significative).

5. Construire alors un intervalle du prix moyen d'un article du rayon "enfant" (choisir un risque $\alpha = 3.6\%$). Quel est le nom de cet intervalle?

On calcule alors un intervalle de CONFIANCE du prix moyen d'un article

$$[\bar{p} - l_{\frac{\alpha}{2}} \sigma / \sqrt{n}; \bar{p} + l_{\frac{\alpha}{2}} \sigma / \sqrt{n}]$$

avec $\bar{p} = 16.2$ et $l_{\frac{\alpha}{2}} = 2.0969$ et on trouve

$$[16.2 - 2.0969 * 7/11; 16.2 + 2.0969 * 7/11] = [14.8656; 17.5344]$$

3) Test dans le cas Bernoulli (variable discrète)

3.1. Test du paramètre d'une loi de Bernoulli de l'hypothèse

$$\underline{H_0: "p = p_0" \text{ contre } H_1: "p \neq p_0"}$$

Si H_0 est vraie alors :

$$\frac{\bar{X} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \xrightarrow{n \text{ grand}} \mathcal{N}(0, 1)$$

Donc,

$$P \left(p_0 - l_{\frac{\alpha}{2}} \sqrt{\frac{p_0(1-p_0)}{n}} \leq \bar{X} \leq p_0 + l_{\frac{\alpha}{2}} \sqrt{\frac{p_0(1-p_0)}{n}} \right) \approx 1 - \alpha$$

D'où le test de $H_0: "p = p_0"$ contre $H_1: "p \neq p_0"$, d'erreur de première espèce (ou niveau) α :

Rejet de H_0 si

$$\bar{x} < p_0 - l_{\frac{\alpha}{2}} \sqrt{\frac{p_0(1-p_0)}{n}} \quad \text{ou} \quad \bar{x} > p_0 + l_{\frac{\alpha}{2}} \sqrt{\frac{p_0(1-p_0)}{n}}$$

C'est un test bilatéral.

3.2. Test du paramètre d'une loi de Bernoulli de l'hypothèse

H_0 : “ $p = p_0$ ” contre H_1 : “ $p > p_0$ ”

Si H_0 est vraie alors :

$$P\left(\frac{\bar{X} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \leq l_\alpha\right) = P\left(\bar{X} \leq p_0 + l_\alpha \sqrt{\frac{p_0(1-p_0)}{n}}\right) \approx 1 - \alpha$$

D'où le test de H_0 : “ $p = p_0$ ” contre H_1 : “ $p > p_0$ ”, d'erreur de première espèce (ou niveau) α :

Rejet de H_0 si

$$\bar{x} > p_0 + l_\alpha \sqrt{\frac{p_0(1-p_0)}{n}}$$

C'est un test unilatéral.

Exemple 1 : Un fabricant de disques compacts affirme qu'au moins 99% de ses disques n'ont aucun défaut. Pour vérifier cette affirmation une association de défense des consommateurs teste 500 disques de ce fabricant et en trouve 10 défectueux. Avec un seuil de 1%, l'association peut-elle contester l'affirmation du fabricant?

Elle doit construire le test de niveau 1%

$$H_0: "p = 0.99" \text{ contre } H_1: "p < 0.99"$$

et vérifier que l'hypothèse H_0 est rejetée.

Le test construit est alors :

Rejet de H_0 si

$$\begin{aligned}\bar{x} &< p_0 - l_\alpha \sqrt{\frac{p_0(1-p_0)}{n}} \\ \bar{x} &< 0.99 - 2.3263 \sqrt{\frac{0.99(1-0.99)}{500}} = 0.9796\end{aligned}$$

où \bar{x} est la proportion de disques “non défectueux”.

Dans l'exemple la proportion de disques “non défectueux” est :

$$\frac{490}{500} = 0.98$$

Exemple 2 : Le même fabricant de disques compacts affirme toujours qu'au moins 99% de ses disques n'ont aucun défaut. Il cherche à conquérir un nouveau marché à l'aide de cette affirmation et doit donc rassurer son nouveau client.

Il lui propose de venir contrôler sa chaîne de production en relevant 1000 ou 2000 exemplaires. Quel sera la démarche de ce nouveau client?

Il va construire le test de niveau α plutôt petit (ex 1%)

$$H_0: "p = 0.99" \text{ contre } H_1: "p > 0.99"$$

et vérifier que l'hypothèse H_0 est rejetée.

Le test construit est alors :

Rejet de H_0 si

$$\begin{aligned}\bar{x} &> p_0 + l_\alpha \sqrt{\frac{p_0(1-p_0)}{n}} \\ \bar{x} &> 0.99 + 2.3263 \sqrt{\frac{0.99(1-0.99)}{2000}} = 0.9952 \\ & \quad \quad \quad (= 0.9973 \text{ si } n = 1000)\end{aligned}$$

où \bar{x} est la proportion de disques “non défectueux”.

Sur les 2000 (resp. 1000) disques prélevés, le fabricant fait donc “le pari” que le nombre de “non défectueux” atteindra au moins :

$0.9952 \times 2000 = 1991$ (resp. 998) éléments.