Master Eseec
Statistique pour l'expertise - partie2

Christian Lavergne

Université Paul Valéry - Montpellier 3
http://moodle-miap.univ-montp3.fr
http://www.univ-montp3.fr/miap/ens

Lois limites de la Statistique et Estimation :

1. Loi des grands nombres et estimation ponctuelle
2. Théorème central limite et estimation par intervalle

La loi de bernoulli et le sondage

• Soit X une expérience aléatoire à 2 états (codés 1/0) : $X \sim \text{Ber}(p)$

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

son espérance mathématique $E(X) = p$
sa variance $V(X) = E(X - E(X))^2 = p(1 - p)$
son écart-type $\sqrt{p(1 - p)}$

• La même expérience (de Bernouilli) répétée n fois de façon indépendante $(X_1, X_2, \ldots, X_n)$ alors

l'espérance mathématique de $\sum_{i}^{n} X_i$ est np ; sa variance est $np(1 - p)$

L'espérance de la somme est la somme des espérances ;
la variance de la somme est la somme des variances

La moyenne de loi de Bernoulli :

- l'espérance mathématique de la moyenne $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ de ces n répétitions est :

$$E(\overline{X}) = E(\frac{1}{n}\sum_{i}^{n} X_i) = p\ .$$

- sa variance

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

- son écart-type

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

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 $X_1, X_2, \ldots, X_n$ 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

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \quad \overset{n \text{ grand}}{\longrightarrow} \quad p$$

Loi des grands nombres et estimation ponctuelle

Soient $X_1, X_2, \ldots, X_n$ n répétitions indépendantes d'une même expérience aléatoire X telle que $E(X) = \mu$ alors :

La moyenne des observations est aussi proche que possible
de la vraie valeur μ
à condition que n soit grand

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \quad \overset{n \text{ grand}}{\longrightarrow} \quad \mu$$

Exemple 1 : Un jeu de la Française des jeux.

Exemple 2 : 2 personnes A et B jouent au dé (à 6 faces) et on propose les gains suivants :

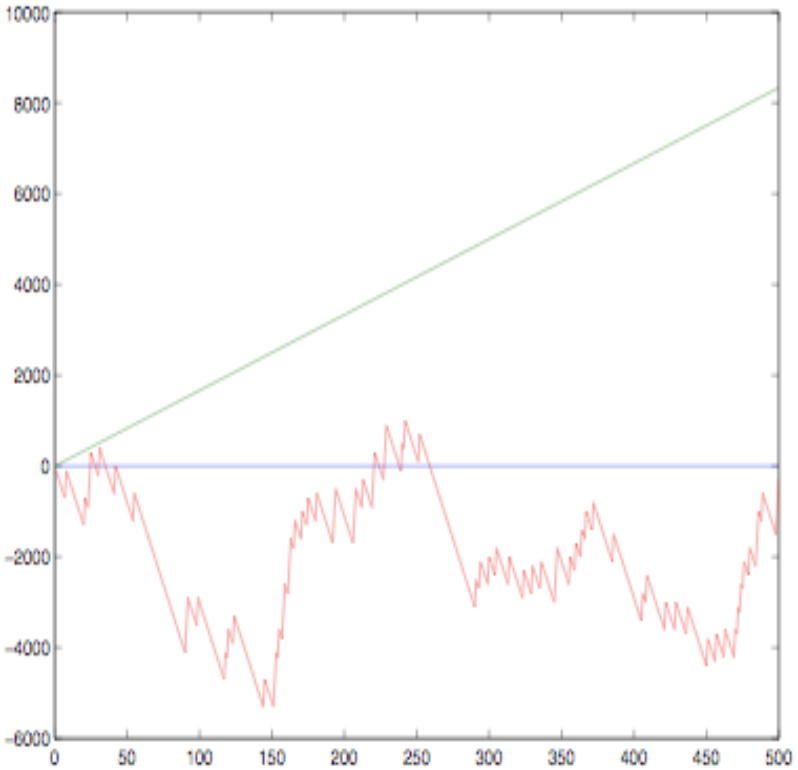
Si "le résultat est la face 6", B paie 600 euros à A
sinon A paie 100 euros à B.

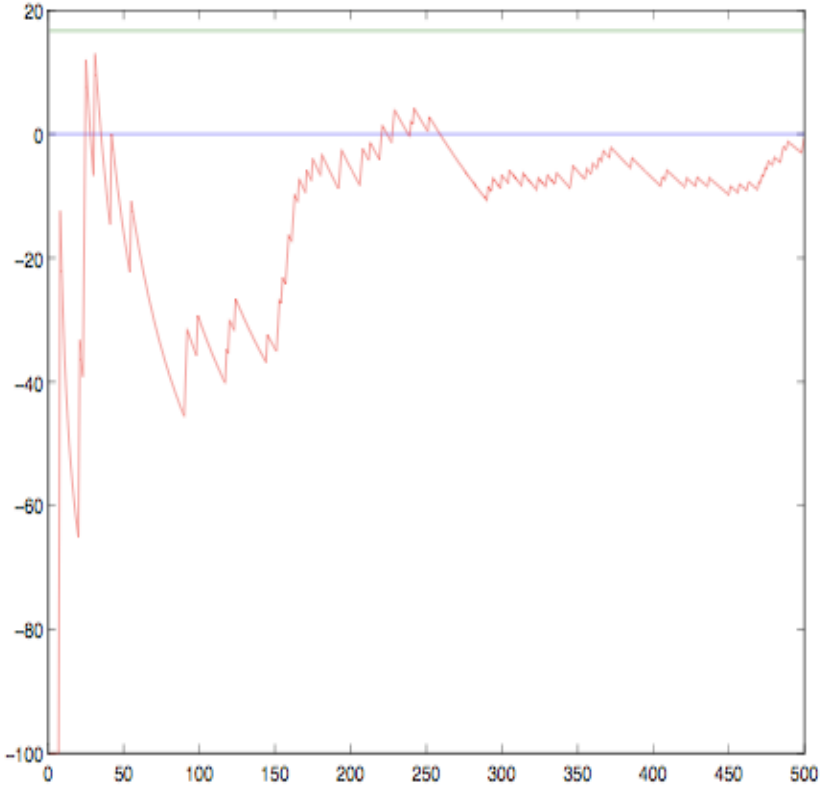
À chaque coup, quelle est l'espérance de gain de A ?
P(A gagne 600) $= \frac{1}{6}$ et P(A perd 100) $= \frac{5}{6}$
donc A a pour espérance de gain :

$$600 \times \frac{1}{6} - 100 \times \frac{5}{6} = \frac{100}{6} = 16.67 \ \textit{euros}$$

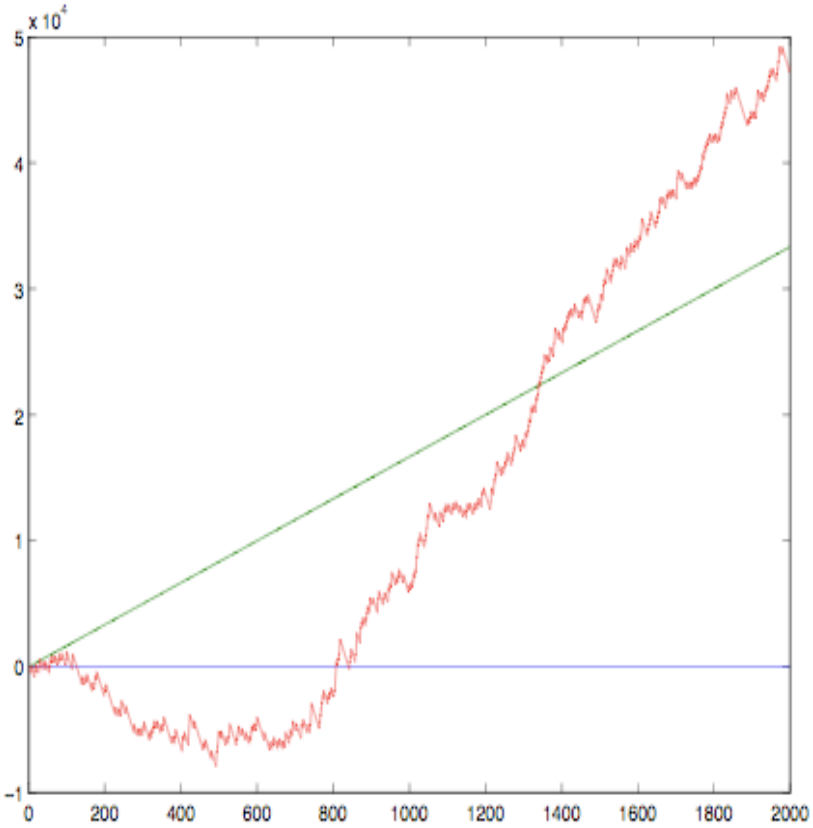
À chaque coup, A gagne 600 ou perd 100, et son espérance de gain est de 16.67.
Donc si A joue un très grand nombre de fois, A est sûr de gagner.

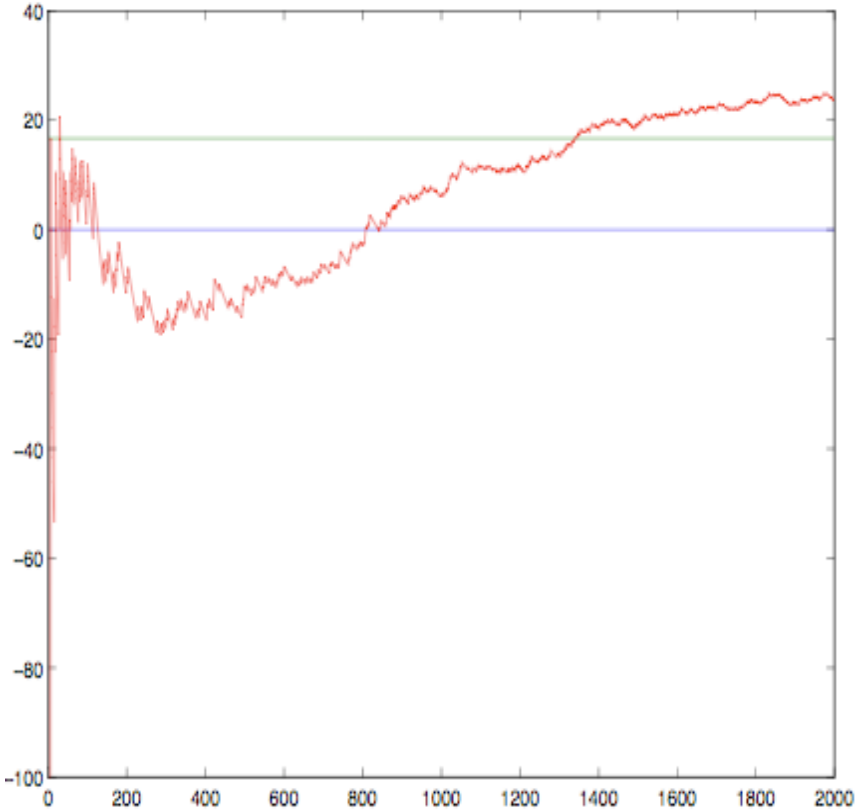
La loi des grands nombres : jeu du dé





La loi des grands nombres : jeu du dé





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

de n répétitions indépendantes $X_1, X_2, \ldots, X_n$ d'une même expérience aléatoire X telle que $E(X) = \mu$ et $V(X) = \sigma^2$:

- son espérance mathématique

$$E(\overline{X}) = \mu$$

- sa variance

$$V(\overline{X}) = \frac{\sigma^2}{n}$$

- son écart-type

$$\sigma(\overline{X}) = \frac{\sigma}{\sqrt{n}}$$

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

Estimation ponctuelle

La moyenne $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ [1] est donc un bon prétendant pour approcher le paramètre inconnu μ.

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

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

Forme générale et propriétés

Soient $X_1, X_2, \ldots, X_n$ n répétitions d'une expérience aléatoire et T_n une combinaison (ou fonction) de ces répétitions.
T_n sera un bon prétendant pour approcher un paramètre inconnu θ ; donc un Estimateur raisonnable de θ ; (T_n sera un $\hat{\theta}$) si :

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

$$E(T_n) = \theta$$

  et on dira que T_n est un estimateur sans biais de θ ;
  mais on peut se contenter de $E(T_n) \overset{n \text{ grand}}{\to} \theta$
- La variance de l'estimateur diminue avec le nombre de répétitions :

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

Exemple : dans la cas de n répétitions indépendantes $X_1, X_2, \ldots, X_n$ d'une même expérience aléatoire X telle que $E(X) = \mu$ et $V(X) = \sigma^2$.

- Si μ est inconnu ; la moyenne
$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \text{ est un estimateur sans biais de } \mu \text{ ; } \hat{\mu}.$$
- Si μ est connu et σ^2 inconnu ; la moyenne des dispersions
$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 \text{ est un estimateur sans biais de } \sigma^2 \text{ ; noté } \hat{\sigma}^2_{\mu}.$$
- Si μ est inconnu et σ^2 inconnu ; la variance empirique
$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 \text{ est un estimateur biaisé de } \sigma^2 \text{ ; noté } \hat{\sigma}^2.$$
- Si μ est inconnu et σ^2 inconnu ;
$$\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \text{ est un estimateur sans biais de } \sigma^2 \text{ ; noté } \hat{\sigma}^2_{SB} \text{ .}$$

Exercice 1 : On suppose avoir observé n variables aléatoires $Y_1, Y_2, \ldots Y_n$ et on propose 3 estimateurs d'un paramètre θ : T_1, T_2, T_3 ayant les propriétés suivants (λ est un autre paramètre inconnu) :

$$E(T_1) = \theta + \lambda \quad \text{et} \quad V(T_1) = (\theta * \lambda)/n$$

$$E(T_2) = \theta + \lambda/n \quad \text{et} \quad V(T_2) = (\theta * \lambda)/n$$

$$E(T_3) = \theta \quad \text{et} \quad V(T_3) = \lambda$$

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

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

• $\overline{X}$, la proportion ou la fréquence de 1 est un Estimateur du paramètre p.

• Après avoir effectué le sondage, la valeur prise par $\overline{X}$ et notée $\overline{x}$ 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.

Théorème central limite et estimation par intervalle

Variable centrée, réduite : *définition*

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

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 alors

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

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

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

Théorème central limite

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

La moyenne se comporte comme une loi normale
à condition que n soit grand
donc centrée et réduite se comporte comme une $\mathcal{N}$ (0,1)

$$\frac{\overline{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \quad \overset{n \text{ grand}}{\rightarrow} \quad \mathcal{N}(0,1)$$

Le sondage

Soient $X_1, X_2, \ldots, X_n$ 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 centrée et réduite se comporte comme une $\mathcal{N}$ (0,1)

$$\frac{\overline{X} - p}{\sqrt{\frac{p(1-p)}{n}}} \overset{n \text{ grand}}{\rightarrow} \mathcal{N}(0,1)$$

Notion élémentaire d'intervalle de dispersion

Terminologie de base : on cherchera à construire un intervalle de la réalisation d'une expérience qui soit d'une grande probabilité.
Étant donné une faible probabilité α, appelé "niveau" (en pratique α vaut 1%, 5% parfois 10%, 0.1%.

Intervalle de dispersion à 1-α

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

$$\mathrm{P}(\mathrm{Nfaces} \in [B_1, B_2]) = 95\%$$

Est ce [220 , 280] ? ou alors [240 , 260] ?

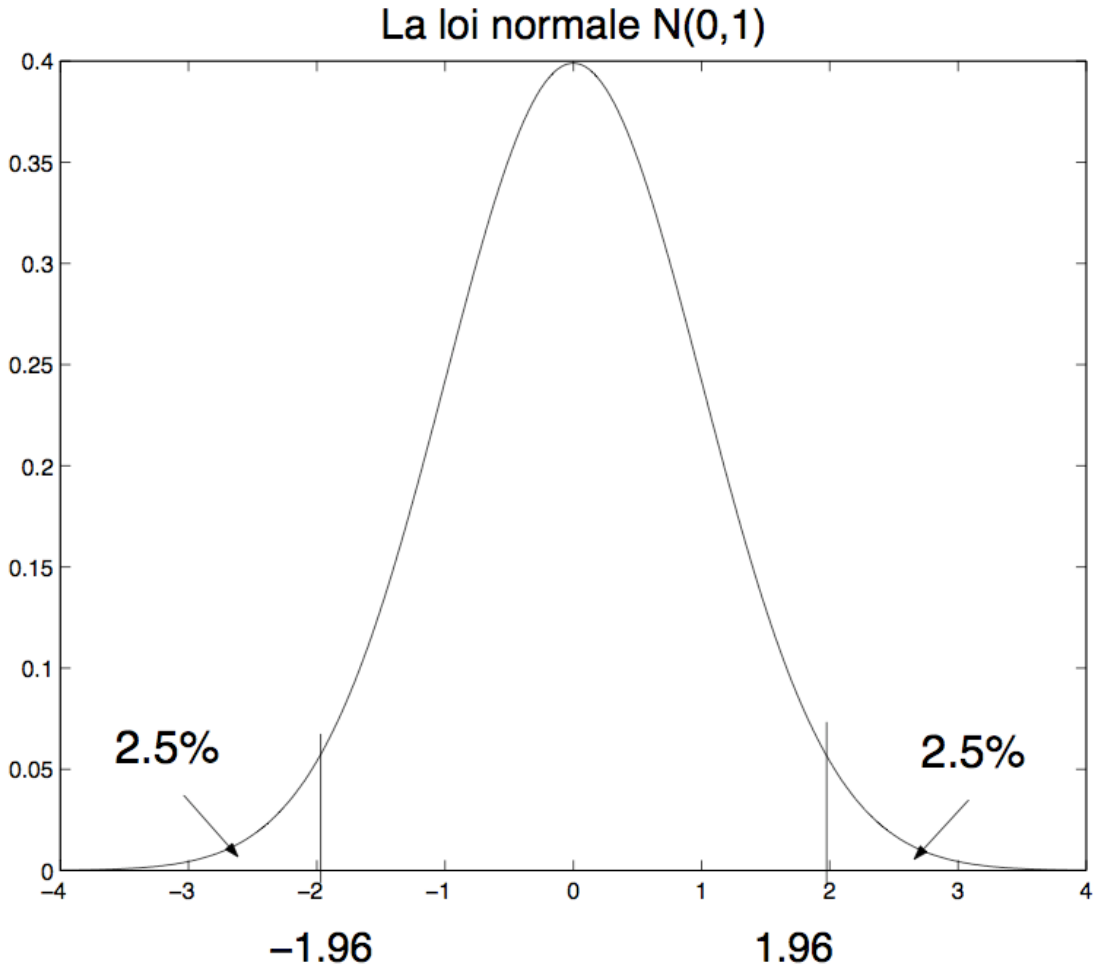
Exemple 2 : Courbes dans les carnets de santé !

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.
À quoi peut-elle s'attendre après la vente de 20 000 produits ?

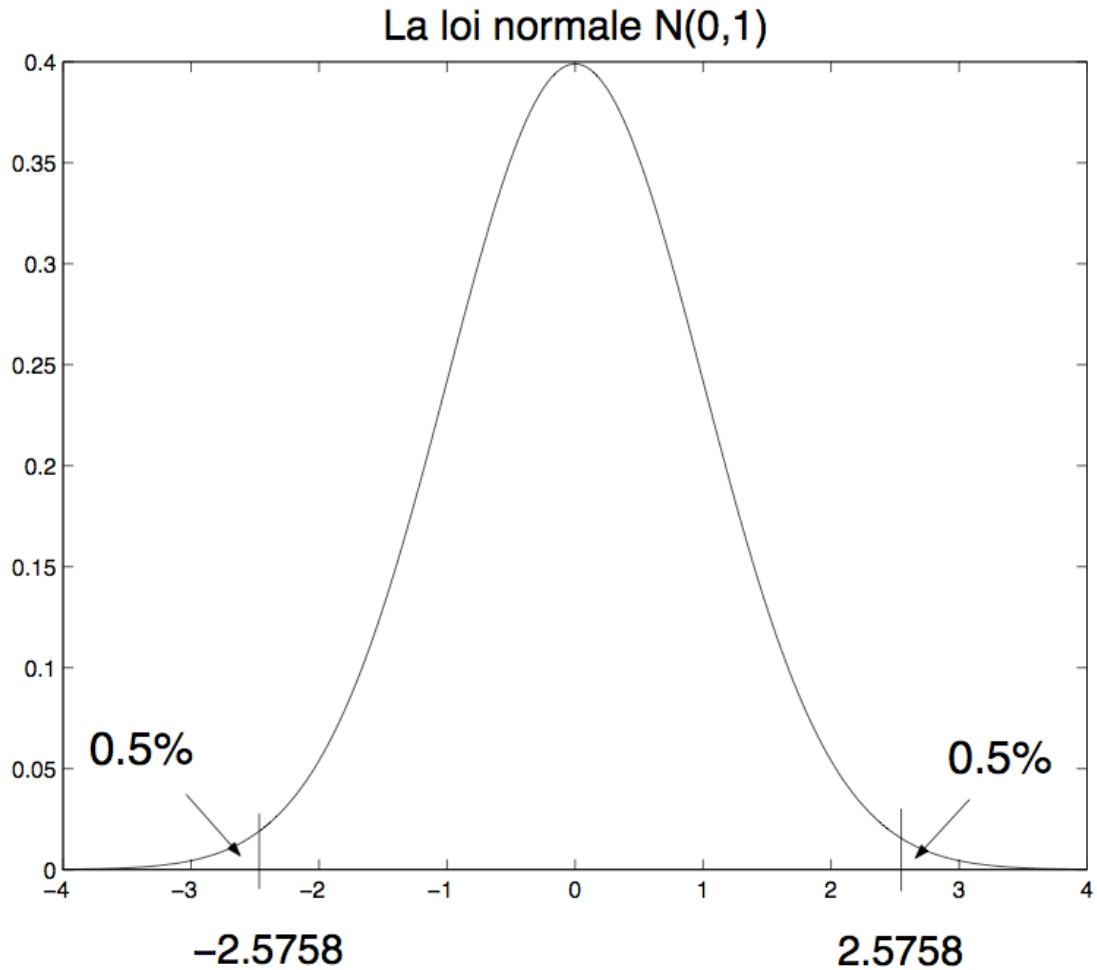
Intervalle de dispersion de la loi normale

Soit Z une variable aléatoire de loi normale N(0,1) alors :



$$P(-1.96 < Z < 1.96) = .95$$

C'est un intervalle de dispersion (ID) à 95%



$$P(-2.5758 < Z < 2.5758) = .99$$

C'est un intervalle de dispersion (ID) à 99%

De façon générale on notera :

$$\left[-I_{\frac{\alpha}{2}} \; ; \; I_{\frac{\alpha}{2}}\right]$$

un intervalle de dispersion (ID) à $1-\alpha$ d'une loi normale $\mathcal{N}(0,1)$

$$P\left(-I_{\frac{\alpha}{2}} < Z < I_{\frac{\alpha}{2}}\right) = 1-\alpha$$

Intervalle de dispersion d'une somme de Bernoulli

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

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

Et on sait si Z est une variable aléatoire N(0,1)

$$P(-I_{\frac{\alpha}{2}} < Z < I_{\frac{\alpha}{2}}) = 1 - \alpha$$

Donc

$$P(-I_{\frac{\alpha}{2}} < \frac{\sum_{i=1}^{n} X_i - np_0}{\sqrt{np_0(1-p_0)}} < I_{\frac{\alpha}{2}}) \approx 1 - \alpha$$

$$\text{Et, } P\left(np_0 - l_{\frac{\alpha}{2}}\sqrt{np_0(1-p_0)} < \sum_{i=1}^{n} X_i < np_0 + l_{\frac{\alpha}{2}}\sqrt{np_0(1-p_0)} \right)$$

$$\approx 1-\alpha$$

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

est donc un **I.D.** de $\sum_{i=1}^{n} X_i$ approximativement à $(1-\alpha)\%$

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

$$\left[500\times\frac{1}{2} - 1.96\sqrt{500\times\frac{1}{2}\times(1-\frac{1}{2})}\ ,\ 500\times\frac{1}{2} + 1.96\sqrt{500\times\frac{1}{2}\times\frac{1}{2}} \right]$$

$$= [250 - 21.91\ ,\ 250 + 21.91] \approx [228\ ,\ 272]$$

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

Notion élémentaire d'intervalle de confiance

Terminologie de base : on cherche à construire un intervalle d'une grandeur particulière inconnue (en général le paramètre d'intérêt d'une loi de probabilité) qui soit d'une grande probabilité.
Étant donné une faible probabilité α, appelée "risque" (en pratique α vaut 1%, 5% parfois 10%, 0.1%.

L'intervalle de confiance à 1-α

que l'on notera : $IC_{1-\alpha}$(paramètre)
représentera un intervalle (dont les bornes seront aléatoires) qui aura 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.

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

les bornes B_1, B_2 de l'intervalle seront calculées à l'aide d'observations.

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

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

Intervalle de confiance du paramètre p

d'une loi de Bernoulli

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

$$\frac{\overline{X} - p}{\sqrt{\frac{p(1-p)}{n}}} \overset{n \text{ grand}}{\rightarrow} \mathcal{N}(0,1)$$

Et on sait que si Z est une variable aléatoire N(0,1)

$$P(-I_{\frac{\alpha}{2}} < Z < I_{\frac{\alpha}{2}}) = 1 - \alpha$$

Donc

$$P(-I_{\frac{\alpha}{2}} < \frac{\overline{X} - p}{\sqrt{\frac{p(1-p)}{n}}} < I_{\frac{\alpha}{2}}) \approx 1 - \alpha$$

Et,

$$P\left(\overline{X} - I_{\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}} < p < \overline{X} + I_{\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}\right)$$

$$\approx$$

$$P\left(\overline{X} - I_{\frac{\alpha}{2}}\sqrt{\frac{\overline{X}(1-\overline{X})}{n}} < p < \overline{X} + I_{\frac{\alpha}{2}}\sqrt{\frac{\overline{X}(1-\overline{X})}{n}}\right)$$

$$\approx 1-\alpha$$

D'où

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

est donc un intervalle de confiance de p approximativement à $(1-\alpha)\%$

Exemple :

$n = 1000$

$\overline{x}$	$\sqrt{\frac{\overline{x}(1-\overline{x})}{n}}$	IC à 95%	IC à 99%
.2	.01265	[0.1752 , 0,2248]	[0.1674 , 0.2326]
.5	.01581	[0.4690 , 0.5310]	[0.4593 , 0.5407]
.9	.00949	[0.8814 , 0.9186]	[0.8756 , 0.9244]

$n = 400$

$\overline{x}$	$\sqrt{\frac{\overline{x}(1-\overline{x})}{n}}$	IC à 95%	IC à 99%
.3	0.0229		
.4			

Intervalle de confiance du paramètre μ

d'une loi gaussienne $\mathcal{N}(\mu, \sigma^2)$

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

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

Et on sait que si Z est une variable aléatoire N(0,1)

$$P(-I_{\frac{\alpha}{2}} < Z < I_{\frac{\alpha}{2}}) = 1 - \alpha$$

Donc

$$P(-I_{\frac{\alpha}{2}} < \frac{\overline{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}} < I_{\frac{\alpha}{2}}) = 1 - \alpha$$

Et

$$P\left(\overline{X} - I_{\frac{\alpha}{2}}\sqrt{\frac{\sigma^2}{n}} < \mu < \overline{X} + I_{\frac{\alpha}{2}}\sqrt{\frac{\sigma^2}{n}}\right) = 1 - \alpha$$

Donc

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

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

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

Ce qui nécessite un grand nombre d'observations n.

Le test statistique :

1. Introduction et Vocabulaire
2. Un test dans le cas continu
3. Un test dans le cas du sondage

Notion élémentaire du 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 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 fabriquant 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 fabriquant ?
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"*.

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(de croire à une "augmentation de la popularité" alors qu'il n'en est rien)

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

Exemple 3 : P(penser à une amélioration dû au groupe alors qu'il n'en est rien)

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 $\mathcal{N}(\mu_0 = 100, \sigma^2)$. 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 $\mathcal{N}(\mu_{groupe}, \sigma^2)$; 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 $\overline{X}$ est gaussienne $\mathcal{N}(\mu, \frac{\sigma^2}{n})$

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

On se place alors sous H_0 vraie ("$\mu = \mu_0$") alors :

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

Donc,

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

Donc 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*

$$\overline{x} > \mu_0 + l_\alpha \frac{\sigma}{\sqrt{n}}$$

C'est un test unilatéral.
Mais le fait de connaître la moyenne $\overline{x}$ ne permet aucune conclusion, la connaissance de la valeur de σ est fondamentale.

Exemple 3 : $\mu_0 = 100$; $n = 80$; pour $\alpha = 5\%$; $l_\alpha = 1.65$
Rejet de H_0 si $\overline{X} > 100 + .184 \times \sigma$.
Donc si $\sigma = 40$, on rejettera si $\overline{x} > 107.38$ et si $\sigma = 80$, on rejettera si $\overline{x} > 114.76$.
Dans l'énoncé, il est donné que $\overline{x}_{groupe} = 110$, on 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 $\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$.

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

$$\widehat{\sigma^2} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

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.

$$\frac{\overline{X} - \mu_0}{\sqrt{\frac{\widehat{\sigma^2}}{n}}} \sim \mathcal{S}_{n-1}$$

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

Test du paramètre d'espérance d'une loi gaussienne

de l'hypothèse H_0 "$\mu = \mu_0$" contre H_1 "$\mu \neq \mu_0$"

Si H_0 est vraie alors :

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

Donc,

$$P\left(\mu_0 - I_{\frac{\alpha}{2}}\sqrt{\frac{\sigma^2}{n}} < \overline{X} < \mu_0 + I_{\frac{\alpha}{2}}\sqrt{\frac{\sigma^2}{n}}\right)$$

$$= 1 - \alpha$$

Donc 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

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

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

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 $\frac{\bar{P}-18}{\sigma/\sqrt{n}}$ est N(0,1)
et on rejettera H_0 si

$$\bar{P} < 18 - I_\alpha \, \sigma/\sqrt{n}$$

avec $I_\alpha = 2.3263$

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 zone de rejet est donc :

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

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 H_0 (le prix moyen ne baisse pas de façon significative) et on proposera alors à la direction d'accentuer sa 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 du rayon "enfant" :

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

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

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

Test du paramètre d'une loi de Bernoulli

de l'hypothèse H_0 "$p = p_0$" contre H_1 "$p \neq p_0$"

Si H_0 est vraie alors :

$$\frac{\overline{X} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \overset{n \text{ grand}}{\rightarrow} \mathcal{N}(0,1)$$

Donc,

$$P\left(p_0 - I_{\frac{\alpha}{2}}\sqrt{\frac{p_0(1-p_0)}{n}} < \overline{X} < p_0 + I_{\frac{\alpha}{2}}\sqrt{\frac{p_0(1-p_0)}{n}}\right)$$

$$\approx 1 - \alpha$$

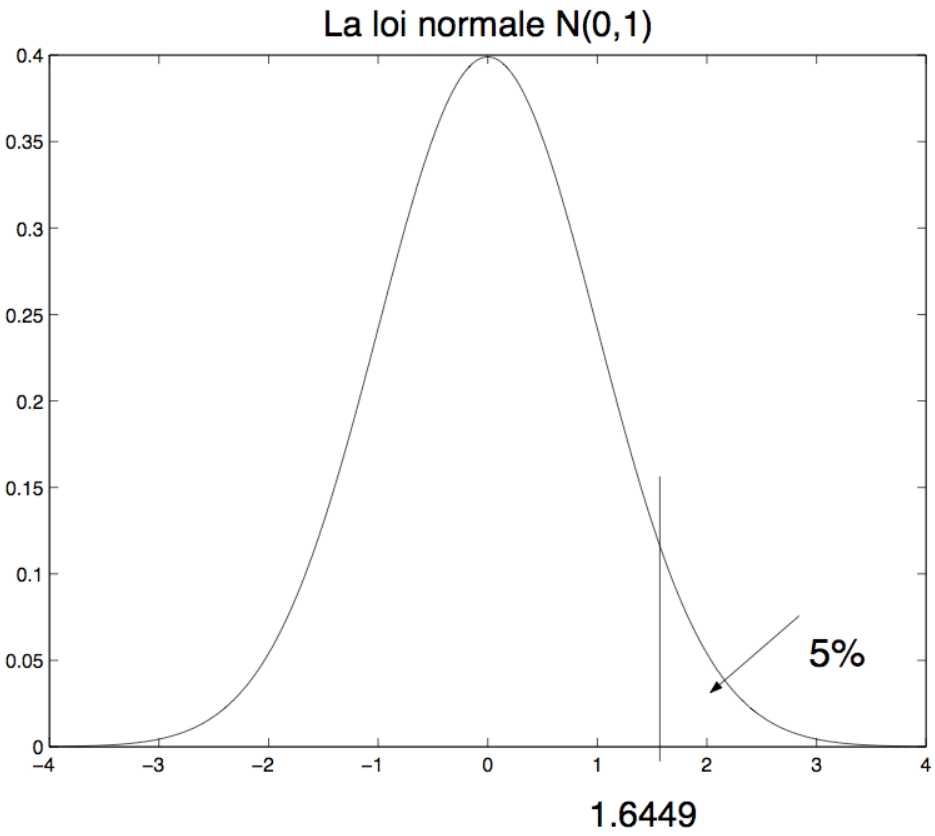
Donc le test de H_0 "$p = p_0$" contre H_1 "$p \neq p_0$", d'erreur de première espèce (ou niveau) α :
Rejet de H_0 si

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

C'est un test bilatéral.

Test du paramètre d'une loi de Bernoulli

de H_0 "$p = p_0$" contre H_1 "$p > p_0$"



$P(Z < 1.6449) = .95$

De façon générale : $P(Z < l_\alpha) = 1 - \alpha$

Si H_0 est vraie alors :

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

Donc 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

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

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 fabriquant 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%

$$H_0 \text{ “}p = 0.99\text{”} \quad \text{contre} \quad H_1 \text{ “}p < 0.99\text{”}$$

et vérifier que l'hypothèse H_0 est rejetée.

Le test construit est alors :
Rejet de H_0 si

$$\overline{x} < p_0 - l_\alpha \sqrt{\frac{p_0(1-p_0)}{n}}$$

$$\overline{x} < .99 - 2.3263 \sqrt{\frac{.99(1-.99)}{500}} = 0.9796$$

où $\overline{X}$ est la proportion de disques “non défectueux".
Dans l'exemple :
la proportion de disques “non défectueux" est

$$\frac{490}{500} = .98$$

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%)

$$H_0 \text{ “}p = 0.99\text{”} \quad \text{contre} \quad H_1 \text{ “}p > 0.99\text{”}$$

et vérifier que l'hypothèse H_0 est rejetée.

Le test construit est alors :
Rejet de H_0 si

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

$$\overline{X} > .99 + 2.3263\sqrt{\frac{.99(1-.99)}{2000}} = 0.9952$$

$$(= 0.9973 \text{ si } n = 1000)$$

où $\overline{x}$ est la proportion de disques “non défectueux".
Sur les 2000 (resp. 1000) disques prélevés, le fabriquant fait donc "le pari" que le nombre de “non défectueux" atteindra au moins :
$.9952 \times 2000 = 1991$ (resp. 998) éléments.