Chapitre II
Estimations

1) Echantillon

- On dispose d'une population Ω et d'une variable aléatoire X sur Ω. (associée à une probabilité $\mathbb{P}$).
- Un échantillon de taille n est la donnée de n éléments de Ω choisis aléatoirement.
- L'échantillon est dit non-exhaustif lorsque le choix / tirage des éléments se fait avec remise.
- Si la taille de Ω est grande par rapport à la taille de l'échantillon, on peut considérer que le tirage des éléments est non exhaustif.
- Soit $E = (\omega_1, \dots, \omega_n)$ un échantillon de taille n. On applique X à chaque individu ω_i pour obtenir un n-uplet de nombres réels :

$$(X(\omega_1), \dots, X(\omega_n))$$

On peut voir les choses de la façon suivante : (2)

$\boxed{\text{On part d'un n-uplet } (X_1, \dots, X_n) \text{ de v.a.i qui suivent la même loi } X.}$

Le n-uplet $(X_1, \dots, X_n)$ est appelé échantillon de X.

Soit $E = (\omega_1, \dots, \omega_n)$ un échantillon de Ω

Posons $x_1 = X_1(\omega_1), \dots, x_n = X_n(\omega_n)$.

On dit que $(x_1, \dots, x_n)$ est une réalisation de $(X_1, \dots, X_n)$.

Définition On appelle statistique de $(X_1, \dots, X_n)$ une variable aléatoire $\varphi(X_1, \dots, X_n)$, où $\varphi : \mathbb{R}^n \to \mathbb{R}$.

Exemples. $\dfrac{X_1 + \dots + X_n}{n}$: moyenne empirique notée $\overline{X}$

- $S^2 = (X_1 - \overline{X})^2 + \dots + (X_n - \overline{X})^2$

$\dfrac{1}{n}\left((X_1 - \overline{X})^2 + \dots + (X_n - \overline{X})^2\right)$ variance empirique notée $\dfrac{1}{n} S^2$.

Une estimation de la statistique $\varphi(X_1, \dots, X_n)$ est la valeur prise par celle-ci sur un échantillon $(\omega_1, \dots, \omega_n)$ de taille n.

On pourra parler de la moyenne d'un échantillon, d'un écart-type d'un échantillon, etc.

<u>Objectif.</u> Comparer, par exemple, la moyenne d'un échantillon avec la moyenne μ_X de la v.a. X

<u>ou encore</u> Peut-on avoir des informations sur μ_X à partir de la moyenne d'un échantillon ?

2) Estimations ponctuelles

Soit X une v.a. de moyenne μ et d'écart-type σ.

Soit la statistique $\overline{X} = \dfrac{X_1 + \dots + X_n}{n}$ (moyenne empirique)

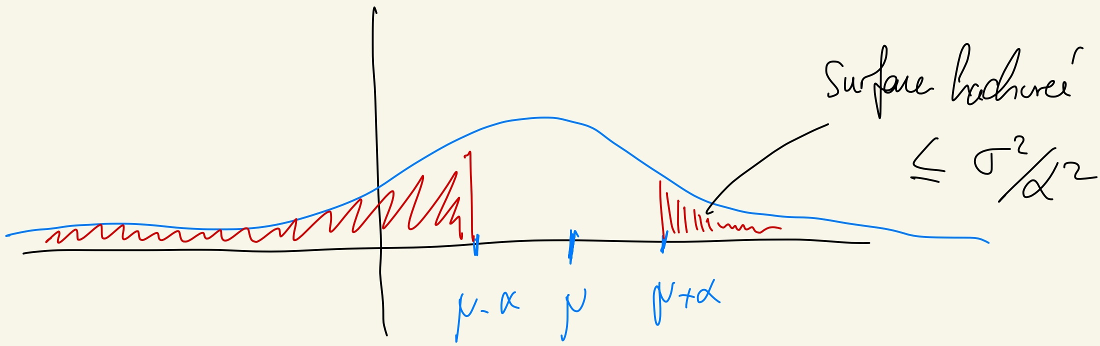
Alors $\boxed{\mu_{\overline{X}} = \mu \text{ et } \sigma_{\overline{X}} = \dfrac{\sigma}{\sqrt{n}}}$ (faits admis)

Inégalité de Bienaymé–Chebyshev

Soit Z une v.a. de moyenne μ et d'écart-type σ.

Alors pour tout nombre $\alpha > 0$:

$$\mathbb{P}(|Z-\mu| \geq \alpha) \leq \sigma^2/\alpha^2$$



Application à $Z = \bar{X} = \dfrac{X_1+\dots+X_n}{n}$.

$$\mathbb{P}(|\bar{X}-\mu| \geq \alpha) \leq \sigma^2/\alpha^2 n$$

C'est le théorème de la loi des grands nombres. Si α est petit et n grand, la probabilité que $\dfrac{x_1+\dots+x_n}{n}$ soit loin de μ, est petite.

⑤

Pour n assez grand, l'inégalité de Bienaymé-Chebychev indique que $\bar{X}(\omega_1, \dots, \omega_n) = \frac{X_1(\omega_1) + \dots + X_n(\omega_n)}{n}$

est une bonne estimation de μ

Ou encore que la moyenne d'un échantillon est une bonne estimation de μ.

Qu'en est-il pour σ^2 ?

Soit $\bar{S}^2 = \frac{1}{n}\left((X_1 - \bar{X})^2 + \dots + (X_n - \bar{X})^2\right)$

Variance empirique.

On a : (fait admis) $\mu_{\bar{S}^2} = \frac{n-1}{n}\sigma^2$ et $\mathrm{Var}(\bar{S}^2)$ petit quand n grand.

Par l'inégalité de Bienaymé-Chebyshev, il vient :

$\bar{S}^2(\omega_1, \dots, \omega_n)$ est une bonne estimation de $\frac{n-1}{n}\sigma^2$

Or pour n grand, $\frac{n-1}{n}\sigma^2$ proche de σ^2.

La variance d'un échantillon est une estimation biaisée de σ^2.

On pose alors :

$$\hat{S}^2 = \frac{n}{n-1}\,\bar{S}^2 = \frac{1}{n-1}\left((X_1-\bar{X})^2 + \dots + (X_n-\bar{X})^2\right)$$

Alors $\boxed{\mu_{\hat{S}^2} = \frac{n}{n-1}\,\mu_{\bar{S}^2} = \sigma^2}$

Par Bienaymé-Chebychev, $\hat{S}^2$ est un bon estimateur de σ^2. (= variance).

Conclusion Soit $(\omega_1, \dots, \omega_n)$ un échantillon de Ω.

Soit $x_1 = X_1(\omega_1), \dots, x_n = X_n(\omega_n)$ une réalisation de $(X_1, \dots, X_n)$.

Soit $\bar{x} = \frac{x_1 + \dots + x_n}{n}$ la moyenne de l'échantillon.

Alors, pour n grand, $\bar{x}$ est une bonne approximation de μ.

et $\frac{1}{n-1}\left((x_1-\bar{x})^2 + \dots + (x_n-\bar{x})^2\right)$ est une bonne approximation de la variance σ^2.

3) Intervalle de confiance pour une moyenne

Soit X une variable aléatoire de moyenne μ et d'écart-type σ.

Problématique

On cherche un intervalle contenant μ, aussi petit que possible, avec une probabilité contrôlée.

L'inégalité de Bienaymé-Chebishev permet d'en avoir un, mais ce n'est pas le "meilleur".

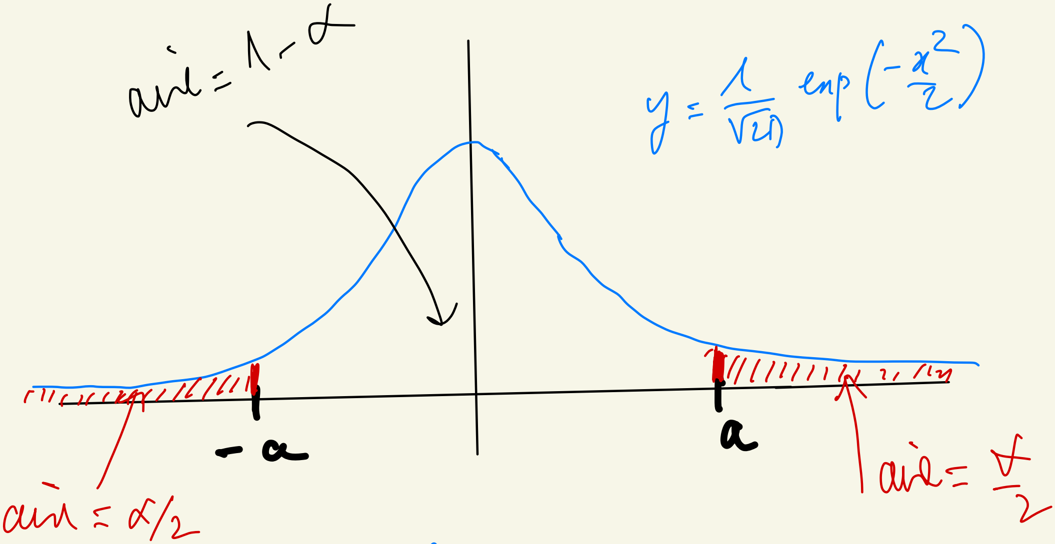
Pour n grand, on utilisera le théorème limite central ... et la courbe de Gauss associée à la loi $\mathcal{N}(0,1)$.

Rappels — Probabilités bilatérales de $\mathcal{N}(0,1)$.

Soit $Y \sim \mathcal{N}(0, 1)$. Soit α donné (erreur).

On cherche a positif tel que :

$$\mathbb{P}(|Y| \geq a) = \alpha$$



On pose $a = t_{\alpha/2}$.

$$\boxed{\begin{aligned} \mathbb{P}(|Y| \geq t_{\alpha/2}) &= \alpha \\ \mathbb{P}(|Y| \leq t_{\alpha/2}) &= 1 - \alpha \end{aligned}}$$

Contexte

- $X_1, \ldots, X_n$ n variables indépendantes qui suivent la même loi X.

- μ = moyenne de X
  σ = écart-type de X.

- $\overline{X} = \dfrac{X_1 + \cdots + X_n}{n}$ $\underset{\text{échantillon}}{\rightsquigarrow}$ $\overline{x} = \dfrac{x_1 + \cdots + x_n}{n}$

On veut donner un intervalle (petit si possible) contenant μ. Il va apparaître 3 cas :

- $X \sim \mathcal{N}(\mu, \sigma)$ avec σ connu

- $X \sim \mathcal{N}(\mu, \sigma)$ avec σ inconnu

- X quelconque, n grand.

Quand $X \sim \mathcal{N}(\mu, \sigma)$ avec σ connu

X_i : n v.a.i avec $X_i \sim \mathcal{N}(\mu, \sigma)$.

$\bar{X} = \frac{X_1 + \dots + X_n}{n}$: moyenne empirique.

Soit $\boxed{Y = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}}$. Alors $Y \sim \mathcal{N}(0,1)$.

(α = erreur)

Ainsi $\mathbb{P}(|Y| \leq t_{\alpha/2}) = 1 - \alpha$.

Il vient ainsi l'intervalle estimé à partir d'un échantillon de taille n :

$$\left| \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \right| \leq t_{\alpha/2}$$

ou encore :

$$\bar{x} - t_{\alpha/2} \sigma / \sqrt{n} \leq \mu \leq \bar{x} + t_{\alpha/2} \sigma / \sqrt{n}$$

$\bar{x}$: moyenne de l'échantillon

μ est dans l'intervalle $\left[\bar{x} - t_{\alpha/2} \sigma / \sqrt{n} \,;\, \bar{x} + t_{\alpha/2} \sigma / \sqrt{n}\right]$ avec une probabilité $1 - \alpha$.

Longueur de l'intervalle : $\frac{2\sigma}{\sqrt{n}} t_{\alpha/2}$: meilleur que B. C.

Exemple. Le caractère d'une population suit une loi normale d'écart-type $\sigma = 2$ mais de moyenne μ inconnue. On dispose d'un taille de taille 100 sur lequel la moyenne des valeurs de X est 2,5. Donner un intervalle de confiance pour μ avec $\alpha = 0{,}05$ (5%)

Ici $t_{\alpha/2} = 1{,}96$.

$$t_{\alpha/2} \times \frac{\sigma}{\sqrt{100}} = 1{,}96 \times \frac{2}{10} = 0{,}392.$$

D'où l'intervalle : $[2{,}5 - 0{,}392 \,;\, 2{,}5 + 3{,}92]$

$$= [2{,}108 \,;\, 2{,}892]$$

Quand $X \sim \mathcal{N}(\mu, \sigma)$ avec σ inconnu

$X_1, \dots, X_n$ n v.a.i. avec $X_i \sim \mathcal{N}(\mu, \sigma)$

$$\bar{X} = \frac{X_1 + \dots + X_n}{n} \quad : \text{moyenne empirique}$$

$$\bar{S}^2 = \frac{1}{n}\left((X_1 - \bar{X})^2 + \dots + (X_n - \bar{X})^2\right) \quad : \text{variance empirique}$$

$$\hat{S}^2 = \frac{1}{n-1}\left((X_1 - \bar{X})^2 + \dots + (X_n - \bar{X})^2\right) = \frac{n}{n-1}\bar{S}^2$$

On rappelle que :

$$\boxed{\frac{\bar{X}-\mu}{\sqrt{\frac{\bar{S}^2}{n-1}}} = \frac{\bar{X}-\mu}{\sqrt{\frac{\hat{S}^2}{n}}} \sim St(n-1)}$$

loi de Student à $n-1$ degrés de liberté

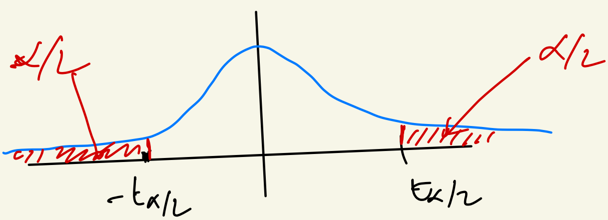
La loi de Student a pour densité une fonction paire.

$\rightsquigarrow$ probabilités bilatérales.

On note $t_{\alpha/2}$ le nombre positif tel que :

$$P(|Y| \geq t_{\alpha/2}) = \alpha$$

Ici $Y \sim St(r)$ où r est donné.



exemple: $n=4$

$r=3$

$\alpha = 0{,}20$ ($=20\%$)

$t_{\alpha/2} = 1{,}64$

Ainsi

$$P\left(\left|\frac{\bar{X}-\mu}{\sqrt{\frac{\hat{S}^2}{n}}}\right| \geq t_{\alpha/2}\right) = \alpha$$

On arrive à l'intervalle (au risque α) :

$$\boxed{\bar{x} - t_{\alpha/2} \hat{\Delta}/\sqrt{n} \leq \mu \leq \bar{x} + t_{\alpha/2} \hat{\Delta}/\sqrt{n}}$$

où
- $\bar{x} = \frac{x_1 + \dots + x_n}{n}$: moyenne de l'échantillon
- $\hat{\Delta}^2 = \frac{1}{n-1}\left((x_1-\bar{x})^2 + \dots + (x_n-\bar{x})^2\right)$
- $\hat{\Delta} = \sqrt{\hat{\Delta}^2}$: estimation de l'écart-type

On remarque que $\sqrt{\frac{n-1}{n}}\,\hat{\Delta}$ = écart-type de l'échantillon.

- $t_{\alpha/2}$ est tel que

$P(|Y| \geq t_{\alpha/2}) = \alpha$ quand $Y \sim St(n-1)$

Le cas d'une loi quelconque X

On utilise le théorème limite central.

X de moyenne μ et d'écart type σ.

$X_1, \ldots, X_n$ n v.a.i avec $X_i \sim X$.

Alors $\overline{X} = \dfrac{X_1 + \cdots + X_n}{n}$ proche de $\mathcal{N}(\mu, \sigma/\sqrt{n})$.

quand **n est grand**.

Comme on prend "n grand", la variance σ^2 sera donnée par l'estimateur :

$$\hat{S}^2 = \frac{1}{n-1}\left((X_1 - \overline{X})^2 + \cdots + (X_n - \overline{X})^2\right).$$

On arrive ainsi à l'approximation suivante :

$$\mathbb{P}\left(\left|\frac{\overline{X} - \mu}{\hat{S}/\sqrt{n}}\right| \leq t_{\alpha/2}\right) = 1 - \alpha$$

donné par la table de la loi $\mathcal{N}(0,1)$

d'où l'intervalle :

$$\boxed{\bar{x} - t_{\alpha/2}\,\hat{s}/\sqrt{n} \leq \mu \leq \bar{x} + t_{\alpha/2}\,\hat{s}/\sqrt{n}}$$

où

- $\bar{x} = \dfrac{x_1 + \cdots + x_n}{n}$ moyenne de l'échantillon

- $\hat{s}^2 = \dfrac{1}{n-1}\left((x_1-\bar{x})^2 + \cdots + (x_n-\bar{x})^2\right)$
  estimation de la variance

- $\hat{s} = \sqrt{\hat{s}^2}$: estimation de l'écart-type

- $t_{\alpha/2}$ est tel que
  $P(|Y| \leq t_{\alpha/2}) = 1-\alpha$
  lorsque $Y \sim \mathcal{N}(0,1)$

4) Intervalle de confiance pour une fréquence (15)

Soit $X \sim \mathcal{B}(p)$ une loi de Bernoulli.

$X_1, \dots, X_n$ des v.a.i. $X_i \sim \mathcal{B}(p)$.

loi binomiale

$\bar{X} = \dfrac{X_1 + \dots + X_n}{n}$: "nombre moyen de succès".

Pour n grand, on peut utiliser le théorème Limite Central.

Pour n grand, $\bar{X}$ est proche de

$$\mathcal{N}\left(p, \frac{\sqrt{p(1-p)}}{\sqrt{n}}\right)$$

c'est à dire

$$\frac{\bar{X} - p}{\sqrt{p(1-p)}} \sqrt{n} \quad \text{proche de} \quad \mathcal{N}(0,1)$$

16

Pour n grand $\mathcal{B}(n,p)$ proche de $\mathcal{N}(np, \sqrt{npq})$

loi binomiale

np et $nq \geq 15$

ou quand

$n > 30$ et np et $nq > 5$.

Pour n grand on obtient l'intervalle de confiance

$$\boxed{\bar{x} - t_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} \leq p \leq \bar{x} + t_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}}$$

donné par $\mathcal{N}(0,1)$

<u>Problème</u>. L'intervalle dépend de p.

- **<u>Une méthode radicale</u>**

<u>Lemme</u>. Soit $0 \leq x \leq 1$. Alors $x(1-x) \leq 1/4$.

Ce qui <u>permet</u> d'obtenir l'intervalle

$$\boxed{\bar{x} - t_{\alpha/2}\frac{1}{2\sqrt{n}} \leq p \leq \bar{x} + t_{\alpha/2}\frac{1}{2\sqrt{n}}}$$

17

- **en estimant la variance** (ici n est grand)

$p(1-p)$ est proche de $\frac{1}{n-1}\left((x_1-\bar{x})^2+\dots+(x_n-\bar{x})^2\right)$

proche de $\frac{1}{n}\left((x_1-\bar{x})^2+\dots+(x_n-\bar{x})^2\right)$

variance de l'échantillon
estimation biaisée.

$$\| $$
$$\bar{x}(1-\bar{x})$$

On obtient l'intervalle :

$$\boxed{\bar{x}-t_{\alpha/2}\sqrt{\frac{\bar{x}(1-\bar{x})}{n}}\leq p \leq \bar{x}+t_{\alpha/2}\sqrt{\frac{\bar{x}(1-\bar{x})}{n}}}$$

où
- $\bar{x}$ = moyenne de l'échantillon
- $\bar{x}(1-\bar{x})$ = variance de l'échantillon
- $t_{\alpha/2}$ est donné par le tableau de $\mathcal{N}(0,1)$

Exercice Lors d'une élection entre deux candidats, on fait un sondage sur 800 personnes pour "estimer" la proportion **p** d'électeurs qui voteront pour A et **q** pour B. Ici $q = 1-p$.

Le sondage donne A à 48% et B à 52%.

1. Donner un intervalle de confiance pour p (avec $\alpha = 5\% = 0{,}05$)

2. Quelle devrait être la taille de l'échantillon pour avoir un intervalle de longueur 1% ? (avec $\alpha = 5\%$).

---

1. Ω = les électeurs. $\mathcal{B}(p)$: loi de Bernoulli.
$n = 800$

$$\overset{p}{\nearrow} A \qquad \overset{q}{\searrow} B$$

Ici $\alpha = 0{,}05$ et $t_{\alpha/2} = 1{,}96$.

$\bar{x} = 0{,}48$

d'où l'intervalle

$$\bar{x} - t_{\alpha/2}\sqrt{\frac{\bar{x}(1-\bar{x})}{n}} \leq p \leq \bar{x} + t_{\alpha/2}\sqrt{\frac{\bar{x}(1-\bar{x})}{n}}$$

Ici $\quad t_{\alpha/2}\sqrt{\dfrac{\bar{x}(1-\bar{x})}{n}} = 1,96 \times \sqrt{\dfrac{0,48 \times 0,52}{800}} \simeq 0,0348$

$$\boxed{0,4453 \leq p \leq 0,5134}$$

Longueur de l'intervalle $\sim 7\%$

(2) En utilisant le "lemme", on obtient un intervalle de longueur : "méthode radicale"

$$\boxed{t_{\alpha/2} \times \frac{1}{\sqrt{n}}}$$

(ne dépend pas de $\bar{x}$).

On veut $\dfrac{t_{\alpha/2}}{\sqrt{n}} \leq 0,01$,

c'est à dire :

$$\sqrt{n} \geq \frac{t_{\alpha/2}}{0,01} = 100 \times 1,96 = 196$$

d'où $\boxed{n \geq 38416}$.

<u>Remarque</u>. Partant de $X \sim \mathcal{N}(\mu, \sigma)$, il est possible de donner un intervalle de confiance pour σ (que l'on connaisse μ ou non). Cela passe par

$$S^2 = (X_1 - \bar{X})^2 + \dots + (X_n - \bar{X})^2 \quad \text{ou}$$

$$S_0^2 = (X_1 - \mu)^2 + \dots + (X_n - \mu)^2$$

puis la loi du **chi-deux χ^2**.

Mais avec une attention particulière : la fonction de densité du χ^2 n'est pas paire (pas de symétrie).