Introduction aux statistiques

I Définitions

Rappel: Soit $n \in \mathbb{N}^*$, on appelle n-échantillon la donnée de n v.a indépendantes $X_1, X_2 \dots, X_n$, iid de même loi.

Généralement, on considère un phénomène aléatoire et une v.a X associée (le résultat de l'expérience)

Tirer un n-échantillon associé à X c'est réaliser n tirages successifs de X (chaque fois de les mêmes conditions) ou réaliser n expériences identiques en parallèle

ex: une urne contient des boules rouges en proportion $p \in [0,1]$ inconnue.

On tire une boule $X = 1$ si la boule est rouge
0 sinon

Tirer un n-échantillon associé à X c'est effectuer n tirages successifs **avec remise**.

II Les intervalles de confiance

a. Le cas gaussien

Soit X une v.a. gaussienne suivant une loi $\mathcal{N}(\mu, \sigma)$ où σ est connu et μ est supposé inconnu. On cherche à estimer μ.

On réalise un n-échantillon associé à X

* On fixe un risque α ($1-\alpha$ s'appelle le niveau de confiance)
* On lit sur les tables (ou on calcule) la valeur de $\varepsilon > 0$ tq
$P(|\bar{X}_n - \mu| < \varepsilon) = 1 - \alpha$

Conséquence (loi des grands nombres)

Soit $X_1, \dots, X_n, \dots$ une suite de v.a indépendantes et de même loi.

On suppose $\mu = E[X_i]$ et $\sigma = \sqrt{V(X_i)}$ sont finis.

On pose :

$$\forall n \in \mathbb{N}, \quad \bar{X}_n = \frac{X_1 + X_2 + \dots + X_n}{n}$$

Alors, $\forall \varepsilon > 0$, $P(|\bar{X}_n - \mu| \geq \varepsilon) \xrightarrow[n \to +\infty]{} 0$

Demo: On applique B.T à $\bar{X}_n$

Rappel : $E(\bar{X}_n) = \mu$ $\quad V(\bar{X}_n) = \frac{\sigma^2}{n}$

$$P(|\bar{X} - \mu| \geq \varepsilon) \leq \frac{V(\bar{X}_n)}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2} \xrightarrow[n \to +\infty]{} 0$$

Interprétation : Qd n est grand, $\bar{X}_n$ est "proche" de μ.

Autre exemple d'utilisation de Bienaymé-Tchebychev

Une étude statistique a montré que le revenu moyen d'une certaine population est 1200 € / mois et l'écart-type 500 €.

Donner un majorant de la proportion de personnes gagnant plus de 3500 € / mois.

Soit X la v.a. "salaire mensuel"

On cherche à estimer $P(X \geq 3500)$

$$P(X \geq 3500) = P(X - 1200 \geq 2300)$$
$$\leq P(|X - 1200| \geq 2300)$$

on applique B.T

$$P(|X - 1200| \geq 2300) \leq \frac{500^2}{2300^2} \leq 4{,}7\%$$

↳ Au plus 4,7% de la pop gagne plus de 3500 € / mois.

II La loi gaussienne

Il s'agit* toute différente des lois précédemment étudiées puisqu'il s'agit d'une loi à densité.

* d'une loi

a. Définition:

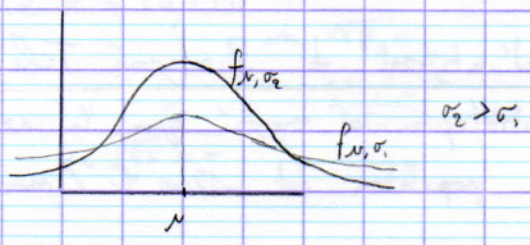
Soit X une v.a prenant ses valeurs dans $\mathbb{R}$ ou entier. On dit que X suit la loi de Gauss (ou normale) de moyenne μ et d'écart-type σ si:

$\forall a < b$, $P(X \in [a,b]) = \int_a^b f_{\mu,\sigma}(x)\,dx$

où: $\forall x \in \mathbb{R}$, $f_{\mu,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

On note: $X \sim N(\mu, \sigma)$

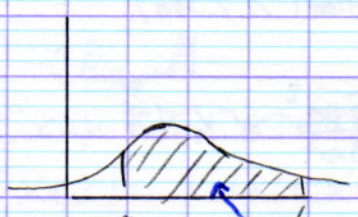
Si $\mu = 0$ et $\sigma = 1$, on dit que X suit la loi normale centrée réduite



La courbe $f_{\mu,\sigma}$ est symétrique par rapport à la droite $[x=\mu]$

Propriété:

- Si $X \sim N(\mu, \sigma)$, X est bien l'espérance de X et σ^2 la variance de X!
- Si $X \sim N(\mu, \sigma)$, $a < b$ $P(X \in [a,b]) = \int_a^b f(x)\,dx$



aire sous la courbe entre a et b
$= P(X \in [a,b])$

+ $\forall a \in \mathbb{R}$, $P(X=a) = \int_a^a f(x)\,dx = 0$

+ $\int_{-\infty}^{+\infty} f_{\mu,\sigma}(x)\,dx = \lim_{x\to+\infty} \int_{-a}^{+a} f_{\mu,\sigma}(x)\,dx = 1$

+ $\forall a < 1$, $P(X \in [a,b]) = P(X \in]a,b[) = P(X \in [a,b[) = P(X \in]a,b[)$
  car $P(X=a) = P(X=b) = 0$

Rem: Supposons $X \sim \mathcal{N}(0,1)$

$P(X \in [0,1]) = \int_0^1 \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}}\,dx$

$= \frac{1}{\sqrt{2\pi}} \int_0^1 e^{-\frac{x^2}{2}}\,dx$

Cette intégrale est incalculable !
↳ on utilise des tables de loi normale

5. Propriétés

Théorème: Soit X une v.a réelle, $\mu \in \mathbb{R}$
$\sigma > 0$ Alors
$X \sim \mathcal{N}(\mu,\sigma) \Leftrightarrow \frac{X-\mu}{\sigma} \sim \mathcal{N}(0,1)$

Démo: Supposons $X \sim \mathcal{N}(\mu,\sigma)$

$P\left(\frac{X-\mu}{\sigma} \in [a,b]\right) = P(X \in [\mu+a\sigma ; \mu+b\sigma])$

$= \int_{\mu+a\sigma}^{\mu+b\sigma} f_{\mu,\sigma}(x)\,dx$

$P\left(\frac{X-\mu}{\sigma} \in [a,b]\right) = \int_{\mu+a\sigma}^{\mu+b\sigma} \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\,dx$

on fait le changement de variable

$t = \frac{x-\mu}{\sigma} \Leftrightarrow x = \mu + \sigma t$

$\Rightarrow dx = \sigma\,dt$

$P\left(\frac{X-\mu}{\sigma} \in [a,b]\right) = \int_a^b \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{t^2}{2}} \sigma\,dt$

$= \int_a^b f_{0,1}(t)\,dt$

$\frac{X-\mu}{\sigma} \sim \mathcal{N}(0,1)$

$\Leftarrow$ idem

Cor : Il suffit de connaître la loi $\mathcal{N}(0,1)$ (cf TD)

Proposition : Si $X \sim \mathcal{N}(\mu, \sigma)$ alors
$\forall \alpha \in \mathbb{R}^*$, $\alpha X \sim \mathcal{N}(\alpha\mu, |\alpha|\sigma)$

Démo :
1er cas : $\alpha > 0$

$$P(\alpha X \in [a,b]) = P\left(X \in \left[\frac{a}{\alpha}, \frac{b}{\alpha}\right]\right)$$
$$= \int_{\frac{a}{\alpha}}^{\frac{b}{\alpha}} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

CV pose $t = \alpha x \Rightarrow dx = dt/\alpha$

$$P(\alpha X \in [a,b]) = \int_a^b \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(\frac{t}{\alpha}-\mu)^2}{2\sigma^2}} \frac{dt}{\alpha}$$
$$= \int_a^b \frac{1}{\sqrt{2\pi}\alpha\sigma} e^{-\frac{(t-\alpha\mu)^2}{2(\alpha\sigma)^2}} dt = \int_a^b f_{\alpha\mu, \alpha\sigma}(t)\, dt$$

2e cas : $\alpha < 0$ (exo)

Théorème (Admis) Soit $\mu_1, \mu_2 \in \mathbb{R}$, $\sigma_1, \sigma_2 > 0$
X et Y 2 v.a indépendantes tq $X \sim \mathcal{N}(\mu_1, \sigma_1)$,
$Y \sim \mathcal{N}(\mu_2, \sigma_2)$
Alors : $X+Y \sim \mathcal{N}(\mu_1+\mu_2, \sqrt{\sigma_1^2 + \sigma_2^2})$

Rem : (Rappel)
$E[X+Y] = E[X] + E[Y] = \mu_1 + \mu_2$
$V(X+Y) = V(X) + V(Y) = \sigma_1^2 + \sigma_2^2$

A ce stade, on fait le pari que l'événement ci-dessus va être réalisé

On fait l'expérience (on réalise l'échantillon), on obtient une valeur $\bar{x}_N$. Comme on a fait le pari que l'event est réalisé, on en déduit que

$$|\bar{x}_N - \mu| < \varepsilon$$

$$|\bar{x}_N - \mu| < \varepsilon \Leftrightarrow \mu \in]\bar{x}_N - \varepsilon, \bar{x}_N + \varepsilon[$$

On conclut : un intervalle de confiance pr μ au risque α est : $]\bar{x}_N - \varepsilon, \bar{x}_N + \varepsilon[$

⚠ μ n'est pas aléatoire

Comment trouver ε ?

$$P(|\bar{x}_N - \mu| < \varepsilon) = P\left(\left|\frac{\bar{x}_N - \mu}{\sigma/\sqrt{N}}\right| < \frac{\varepsilon\sqrt{N}}{\sigma}\right)$$

or $\bar{X}_N \sim \mathcal{N}(\mu, \frac{\sigma}{\sqrt{N}})$

donc $\frac{\bar{x}_N - \mu}{\sigma/\sqrt{N}} \sim \mathcal{N}(0,1)$

On lit sur la table de la loi $\mathcal{N}(0,1)$ la valeur de t tq

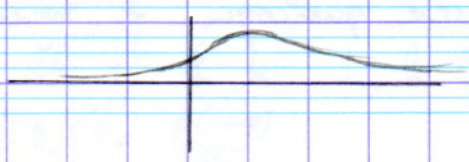
$P(|Z| < t) = 1 - \alpha$ si $Z \sim \mathcal{N}(0,1)$

t est fixé, indépendant de N, μ, X

$$\frac{\varepsilon\sqrt{N}}{\sigma} = t \Leftrightarrow \boxed{\varepsilon = \frac{\sigma t}{\sqrt{N}}}$$

On voit que plus N est grand, plus ε est petit

ex: La durée de vie d'une ampoule électrique suit une loi $\mathcal{N}(\mu, 50)$



On teste 10 ampoules, et on mesure les durées suivantes: 2115, 1955, 2005, 1995, 2100, 2225, 1955, 1990, 2110, 2050.
(les qtés st exprimées en heures)

Donner un intervalle de confiance pr μ au risque 5%
Soit X la v.a. "durée de vie d'une ampoule"
$n = 10$, $X_1, X_2, \dots, X_{10}$ un 10-échantillon
on cherche ε tq $P(|\bar{X}_{10} - \mu| < \varepsilon) = 0,95$
$\bar{X}_{10} \sim N(\mu, \frac{\sigma}{\sqrt{10}})$
$P(|\bar{X}_{10} - \mu| < \varepsilon) = P(|\frac{\bar{X}_{10} - \mu}{\sigma/\sqrt{10}}| < \frac{\varepsilon\sqrt{10}}{\sigma})$

$\frac{\bar{X}_{10} - \mu}{\sigma/\sqrt{10}} \sim N(0,1)$

On lit sur la table de la loi $N(0,1)$
$t = \frac{\varepsilon\sqrt{10}}{\sigma} = 1,96$

$\varepsilon = 1,96 \times \frac{\sigma}{\sqrt{10}} \simeq 31$

Sur l'échantillon, on mesure $\bar{x}_{10} = \frac{2115 + \dots + 2050}{10}$
$= 2033$
On fait le pari que l'event $|\bar{X}_{10} - \mu| < \varepsilon$ est réalisé par l'échantillon
$|2033 - \mu| \leq 31 \Rightarrow \mu \in [2033 - 31, 2033 + 31]$

Un intervalle de confiance pr μ au risque 5% est
$[2002, 2064]$

b. <u>Le cas général</u>
Supposons X une v.a. quelconque de moyenne μ inconnue.

On fait le même raisonnement

On cherche ε tq $P(|\bar{X}_n - \mu| < \varepsilon)$

→ Soit on connait la loi de $\bar{X}_n - \mu$ on fait le calcul

→ Soit n est grand, le TCL assure que $\bar{X}_n - \mu$ est proche d'une loi gaussienne et on est ramenés au cas précédent (si on connait l'écart-type de X)

ex: On considère un dé pipé. On cherche à estimer p, la proba d'obtenir 6. On lance le dé 400 fois et on obtient 147 fois "6". Donner un intervalle de confiance pour p, au risque 1%.

On considère le 400-échantillon

$X_i = \begin{cases} 1 \text{ si on a obtenu 6 au } i\text{-ème lancer} \\ 0 \text{ sinon} \end{cases}$

$X_i \sim \mathcal{B}(p)$ (rappel: p est inconnu)

$\bar{X}_n = \frac{X_1 + \dots + X_{400}}{400}$

$E[\bar{X}_n] = p \quad V(X_i) = p(1-p) \Rightarrow V(\bar{X}_n) = \frac{p(1-p)}{\sqrt{400}} = \frac{p(1-p)}{20}$

Pb: Ici, l'écart type de X est non connu !

On fait l'approximation gaussienne

$\frac{\bar{X}_{400} - p}{\frac{p(1-p)}{20}} \sim \mathcal{N}(0,1)$

On cherche ε tq $P(|\bar{X}_n - p| < \varepsilon) = 0,99$

$P\left(\underbrace{\frac{|\bar{X}_{400} - p|}{\frac{p(1-p)}{20}}}_{\sim \mathcal{N}(0,1)} < \underbrace{\varepsilon \times \frac{20}{p(1-p)}}_{t} \right) = 0,99$

On lit sur la table de la loi $\mathcal{N}(0,1)$ $t = 2,58$

$\varepsilon = 2,58 \times \frac{p(1-p)}{20}$

On suppose l'event $|\bar{x}_{400} - p| < \varepsilon$ réalisé par l'échantillon

$\bar{x}_{400} = \frac{147}{400}$

$|\bar{x}_{400} - p| < \varepsilon \Leftrightarrow p \in [\bar{x}_{400} - \varepsilon, \bar{x}_{400} + \varepsilon]$

$\varepsilon = \frac{2,58}{20} \times p(1-p)$

On vérifie que $\forall p \in [0,1]$, $p(1-p) \leq 1/4$

$\forall p \in [0,1]$, $\varepsilon \leq \frac{2,58}{20} \times \frac{1}{4} = \frac{2,58}{80}$

$\hookrightarrow [\bar{x}_{400} - \varepsilon, \bar{x}_{400} + \varepsilon] \subset [\bar{x}_{10} - \frac{2,58}{80}, \bar{x}_{10} + \frac{2,58}{80}]$

$\hookrightarrow$ Un intervalle de confiance pour p, au risque 1% est $[0,33; 0,4]$

Rem: Dans ce genre de situations, ak de grands effectifs, les statisticiens considèrent que $p(1-p) \simeq \bar{x}_{400}(1 - \bar{x}_{400})$

VI Introduction à la théorie des test

a. Exemple

Une machine produit des tiges d'acier de longueur 1 m, si la machine est bien réglée.

On soupçonne que celle-ci est mal réglée et produit des tiges de longueur inférieure (à 1 m).

On veut, au vu d'un échantillon de la production, décider si la machine est déréglée ou non.

Modélisation: X v.a. "longueur d'une tige"

$X \sim N(\mu, \sigma)$

On simplifie, on suppose que σ est connu $\sigma = 5\,cm = 0,05\,m$.

Choix des hypothèses: Elles sont 2 H_0 : "la machine est bien réglée", $\mu = 1m$

H_1 : "la machine est déréglée", $\mu < 1m$

Fixer un risque: En fait il y a 2 risques

Réalité \ Décision	H_0	H_1
		α
H_0	OK	Risque de 1ère espèce
H_1	β risque de 2ème espèce	OK