Théorème central limite, tests d'hypothèse

1 Théorème central limite et applications

1.1 TCL

Soit $(X_n)_{n\geq 1}$ des v.a. indépendantes de même loi, et $S_n = X_1 + \cdots + X_n$.
On note $m = E(X_1) = \cdots = E(X_n)$ et $\sigma^2 = Var(X_1) = \cdots = Var(X_n)$.

Théorème. La loi de $\dfrac{S_n - nm}{\sqrt{n\sigma^2}}$ converge vers $\mathcal{N}(0,1)$ quand n tend vers l'infini, c'est-à-dire :

$$\forall a, b \in \mathbb{R} \cup \{-\infty, +\infty\} \text{ avec } a \leq b, \ \lim_{n\to+\infty} P\left(a < \frac{S_n - nm}{\sqrt{n\sigma^2}} < b\right) = P(a < N < b),$$

où N suit la loi normale standard $\mathcal{N}(0,1)$, c'est-à-dire $P(a < N < b) = \displaystyle\int_a^b \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$.

Remarque : on est souvent intéressé par $\frac{S_n}{n}$ plutôt que S_n. On a : $\dfrac{S_n - nm}{\sqrt{n\sigma^2}} = \sqrt{n}\left(\dfrac{\frac{S_n}{n} - m}{\sigma}\right)$

Si n est assez grand (on considère souvent $n \geq 100$ comme "assez grand"), on approxime la loi de $\sqrt{n}\left(\frac{\frac{S_n}{n} - m}{\sigma}\right)$ par $\mathcal{N}(0,1)$.

1.2 Intervalles de confiance

En pratique : le TCL est beaucoup utilisé pour des intervalles de confiance.
On choisit un seuil d'erreur ε (par exemple $\varepsilon = 0,05$) ou un niveau de confiance $\alpha = 1 - \varepsilon$. Avec les tables de la loi normale standard, on trouve un intervalle $[a,b]$ tel que $P(a \leq \mathcal{N}(0,1) \leq b) \geq \alpha$. On prend souvent un intervalle symétrique, c'est-à-dire $a = -b$ (parfois on a une raison particulière de prendre un intervalle du type $]-\infty, b]$ ou $[a, +\infty[$). On en déduit que, pour n assez grand,

$$P\left(a < \sqrt{n}\left(\frac{\frac{S_n}{n} - m}{\sigma}\right) < b\right) \geq \alpha.$$

On transforme alors l'encadrement en fonction de ce qu'on cherche. Voici les deux situations les plus fréquentes :
1) La loi des X_i est connue (donc m et σ sont connus) :

$$-b \leq \sqrt{n}\left(\frac{\frac{S_n}{n} - m}{\sigma}\right) \leq b \iff m - \frac{b\sigma}{\sqrt{n}} \leq \frac{S_n}{n} \leq m + \frac{b\sigma}{\sqrt{n}}.$$

Cela fournit un intervalle $I_n = \left[m - \frac{b\sigma}{\sqrt{n}}, m + \frac{b\sigma}{\sqrt{n}}\right]$ contenant $\frac{S_n}{n}$ avec une probabilité au moins α, autrement dit $P\left(\frac{S_n}{n} \in I_n\right) \geq \alpha$

Contexte typique : on veut savoir, avant l'expérience, où se situera le résultat $\frac{S_n}{n}$ la plupart du temps.

2) La loi des X_i est inconnue et on cherche à évaluer m à partir du résultat d'une expérience. Soit on connaît σ, soit on l'estime (avec les données de l'expérience).

$$-b \leq \sqrt{n}\left(\frac{\frac{S_n}{n} - m}{\sigma}\right) \leq b \iff \frac{S_n}{n} - \frac{b\sigma}{\sqrt{n}} \leq m \leq \frac{S_n}{n} + \frac{b\sigma}{\sqrt{n}}$$

Cela fournit un intervalle $I_n = [\frac{S_n}{n} - \frac{b\sigma}{\sqrt{n}}, \frac{S_n}{n} + \frac{b\sigma}{\sqrt{n}}]$ contenant m avec probabilité au moins α, autrement dit $P(m \in I_n) \geq \alpha$ (remarque : m est fixé, c'est I_n qui varie car il dépend d'une variable aléatoire ; l'intervalle prend donc des valeurs différentes suivant l'expérience !)

Pour avoir un intervalle concret, on réalise une expérience, ce qui revient à évaluer les v.a. en ω pour un certain ω choisi au hasard. On obtient des données $x_1(= X_1(\omega)), \ldots, x_n(= X_n(\omega))$ et $\bar{x} = \frac{x_1 + \cdots + x_n}{n} (= \frac{S_n(\omega)}{n})$. L'intervalle de confiance déterminé par l'expérience est alors $I = [\bar{x} - \frac{b\sigma}{\sqrt{n}}, \bar{x} + \frac{b\sigma}{\sqrt{n}}]$.

Exemple typique : on fait un sondage (réponse "oui"=1, "non"=0), on sait que les X_i suivent une loi de Bernoulli de paramètre p (où p est la proportion de personnes dans la population totale pensant "oui") mais on ne connaît pas p. On sait que $m = p$ et $\sigma^2 = p(1-p)$. On ne connaît pas σ^2 (puisque p n'est pas connu), il faut donc l'estimer. La loi des grands nombres dit que $\frac{S_n}{n}$ converge vers p et on estime σ^2 par $\frac{S_n}{n}(1 - \frac{S_n}{n})$, ou plutôt par $\bar{x}(1-\bar{x})$, où $\bar{x} = \frac{S_n(\omega)}{n}$ est obtenu avec les données de l'expérience (remarque : on ne remplace pas carrément m par $\frac{S_n}{n}$ car on cherche m... et on se retrouverait avec l'encadrement $-b \leq 0 \leq b$, qui n'apprend rien !).

1.3 Test sur un paramètre

Soit X une variable aléatoire dont la loi dépend d'un paramètre p inconnu (par exemple, X suit une loi de Bernoulli $b(p)$, ou une loi de Poisson $\mathcal{P}(p)$, ...). On veut comparer p à un paramètre p_0 fixé. On fait deux hypothèses :
– $H_0 : p = p_0$,
– $H_1 : p \neq p_0$ (ou $p > p_0$, ou $p < p_0$, si la situation incite à privilégier l'un ou l'autre cas).
On dit qu'on teste H_0 contre H_1. On prend toujours pour H_0 une égalité, de sorte que, sous l'hypothèse H_0, on connaît la loi de X. On choisit un niveau de test ε (par exemple $\varepsilon = 0,05$), qui est la probabilité de rejeter H_0 (et donc de conclure H_1) alors que H_0 est vraie.
On prend $X_1, \ldots, X_n$ un n-échantillon de même loi que X (c'est-à-dire que $X_1, \ldots, X_n$ sont indépendantes et ont toutes la même loi que X). On se place sous l'hypothèse H_0 et on calcule un intervalle de confiance I_n pour $\frac{S_n}{n}$, où $S_n = X_1 + \cdots + X_n$ (on est dans le cas 1 de la partie 1.2 puisque la loi de X est connue sous H_0). Cet intervalle I_n est la zone d'acceptation du test.
On fait l'expérience, ce qui revient à évaluer les v.a. en ω pour un certain ω choisi au hasard. On obtient des données $x_1(= X_1(\omega)), \ldots, x_n(= X_n(\omega))$ et $\bar{x} = \frac{x_1+\cdots+x_n}{n} (= \frac{S_n(\omega)}{n})$.
– si $\bar{x} \in I_n$, on conclut que H_0 est vraie,
– si $\bar{x} \notin I_n$, on conclut que H_1 est vraie.
On a une probabilité $\leq \varepsilon$ de conclure H_1 alors que c'est H_0 qui est vraie.

Remarque : si $H_1 =$"$p \neq p_0$", on choisit un intervalle symétrique (c'est-à-dire $a = -b$ dans le TCL). Si $H_1 =$"$p > p_0$", on cherche plutôt à majorer $\frac{S_n}{n}$ (s'il est trop grand, c'est louche sous H_0 et on privilégie H_1), donc on prend plutôt $a = -\infty$ et b tel que $P(N \leq b) \geq 1 - \varepsilon$. L'intervalle d'acceptation est alors $]-\infty, m + \frac{b\sigma}{\sqrt{n}}]$ (mais on peut quand même choisir un intervalle symétrique ; ce n'est pas faux, c'est juste un peu moins adapté à la situation). Si $H_1 =$"$p < p_0$", c'est l'inverse.

2 Test d'ajustement du χ_2

Soit X une v.a. prenant ses valeurs dans l'ensemble fini $\{1, \ldots, d\}$. On suppose qu'on connaît la loi de X et on note $p_k = P(X = k)$ pour tout $k \in \{1, \ldots, d\}$. Soit Y une v.a. prenant aussi ses valeurs dans $\{1, \ldots, d\}$, de loi inconnue. On se demande si la loi de Y est la même que la loi de X. On considère $Y_1, \ldots, Y_n$ un n-échantillon de même loi que Y.
Pour tout k dans $\{1, \ldots, d\}$, on pose

$$N_n^{(k)} = \text{nombre d'indices } i \in \{1, \ldots, n\} \text{ tels que } Y_i = k.$$

$N_n^{(k)}$ est une variable aléatoire, $N_n^{(k)}(\omega)$ = nombre de fois que k apparaît dans l'expérience réalisée (le choix de l'expérience correspondant à choisir un ω).
On pose $Z_n = \displaystyle\sum_{k=1}^{d} \frac{(N_n^{(k)} - np_k)^2}{np_k}$.

Théorème. Si la loi de Y est la même que la loi de X, alors la loi de Z_n converge vers $\chi_2(d-1)$ (loi du χ_2 à $d-1$ degrés de liberté) quand n tend vers l'infini. Sinon, Z_n tend vers $+\infty$.

En pratique, on fait les hypothèses :
– H_0 : "Y a la même loi que que X",
– H_1 : "Y n'a pas la même loi que X".
Sous l'hypothèse H_0, on approxime la loi de Z_n par $\chi_2(d-1)$ si n et tous les np_k sont assez grands (on considère souvent comme "assez grands" $n \geq 100$ et $np_k \geq 10$ pour tout k).
On choisit un niveau de test ε, qui est la probabilité de rejeter H_0 à tort. On calcule z_{obs} la valeur observée de Z_n à partir des données (ce qui revient à évaluer les v.a. en un certain ω, qui correspond à l'expérience réalisée, et $z_{obs} = Z_n(\omega)$, calculé à partir des valeurs $y_1(= Y_1(\omega)), \ldots, y_n(= Y_n(\omega))$, qui sont les données mesurées lors de l'expérience) et on regarde dans les tables la valeur de :

$$P(\chi_2(d-1) \geq z_{obs}).$$

Si cette probabilité est plus petite que le niveau de test ε, on rejette l'hypothèse H_0 et on accepte H_1 (c'est-à-dire qu'on conclut que la loi de Y n'est pas la même que celle de X, avec un risque de 5% de se tromper). Si elle est plus grande, on accepte l'hypothèse H_0.