7 Il teorema del limite centrale

La formulazione più semplice del teorema limite centrale che si incontra in un primo corso di probabilità è la seguente.

Theorem 64. *Sia $X_1, ..., X_n$ una seuenza di variabili aleatorie indipendenti ed identicamente distribuite con valor medio μ e varianza (finita) σ^2. Allora abbiamo la convergenza in distribuzione*

$$\frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n\sigma^2}} \to_d Z \sim N(0,1), \text{ per } n \to \infty$$

Spiegazione Semplice

Questo è il Teorema del Limite Centrale (CLT)
Questa è forse la formula più importante di tutta la statistica.
Cosa dice in parole povere? Prendi *qualsiasi* variabile aleatoria X (es. il lancio di un dado, che ha una distribuzione "piatta"). Non importa quale sia la sua forma originale, purché abbia una media μ e una varianza σ^2 finite.
Ora, prendi un campione di n osservazioni $(X_1, \ldots, X_n)$ e calcola la loro **somma** $\sum X_i$.
Il teorema dice che, per n grande, la *distribuzione* di questa somma (o della sua media), una volta "standardizzata", assomiglierà sempre e comunque a una Gaussiana Standard $N(0, 1)$ (la "curva a campana").
Cosa significa "Standardizzare"? È un'operazione di "zoom e centratura" per guardare la distribuzione alla giusta scala:

1. $\sum X_i - n\mu$: Centriamo la somma sottraendo il suo valore medio.
2. $/\sqrt{n\sigma^2}$: Riduciamo la scala (dividendo per la sua deviazione standard) per evitare che la curva si "allarghi" o "schiacci" all'infinito.

Questo teorema ci dà il permesso di usare la curva Normale (Gaussiana) per fare stime e test d'ipotesi su quasi tutto, purché abbiamo abbastanza dati (n grande).

Dimostrazione. Ricordiamo innanzitutto la definizione di funzione caratteristica di una variabile aleatoria X:

$$\psi_X(t) := E[exp(iXt)].$$

Ricordiamo altri due fatti dai corsi elementari di probabilità: Il Teorema di continuità di Lévy ci garantisce che se la sequenza di funzioni di caratteristiche $\psi_{X_n}(t) = E[exp(iX_n t)]$ converge a $\psi_X(t) = E[exp(iXt)]$, necessariamente si ha la convergenza in distribuzione $X_n \to_d X$ (eviceversa). Ovviamente nel nostro caso è sufficiente dimostrare che

$$\frac{\sum_{i=1}^{n} \tilde{X}_i}{\sqrt{n}} \to_d Z \sim N(0,1), \text{ per } n \to \infty, \text{ con } \tilde{X}_i := \frac{X_i - \mu}{\sigma}$$

Abbiamo dunque, usando le proprietà standard della funzione caratteristica

$$\begin{aligned}
\psi_{\frac{\sum_{i=1}^{n} \tilde{X}_i}{\sqrt{n}}}(t) &= E[\exp(i \sum_{i=1}^{n} \tilde{X}_i \frac{t}{\sqrt{n}})] \\
&= \psi_{\sum_{i=1}^{n} \tilde{X}_i}(\frac{t}{\sqrt{n}}) \\
&= \psi_{\tilde{X}_1}(\frac{t}{\sqrt{n}}) \times ... \times \psi_{\tilde{X}_n}(\frac{t}{\sqrt{n}}) \\
&= \{\psi_{\tilde{X}_1}(\frac{t}{\sqrt{n}})\}^n
\end{aligned}$$

dove nel secondo passaggio abbiamo usato il fatto che le variabili sono indipendenti e nel terzo il fatto che sono identicamente distribuite. Ricordiamo ora che

$$\psi_{\tilde{X}_1}(t) = 1 + itE[\tilde{X}_1] + \frac{(it)^2}{2} E[\tilde{X}_1^2] + o((it)^2),$$

da cui

$$\{\psi_{\tilde{X}_1}(\frac{t}{\sqrt{n}})\}^n = \{1 + i\frac{t}{\sqrt{n}}E[\tilde{X}_1] + \frac{(it)^2}{2n}E[\tilde{X}_1^2] + o((\frac{it}{\sqrt{n}})^2)\}^n$$
$$= \{1 - \frac{t^2}{2n} + o(\frac{t^2}{n})\}^n \to \exp(-\frac{t^2}{2}).$$

La dimostrazione è conclusa ricordando che la Gaussiana standard ha proprio funzione caratteristica

$$\psi_Z(t) = \exp(-\frac{t^2}{2}).$$

□

Spiegazione Semplice

Come si dimostra il Teorema (l'idea)?
La dimostrazione formale è tecnica, ma l'idea è elegante e si basa su uno strumento chiamato **Funzione Caratteristica** $\psi_X(t)$.

- **Cos'è?** È una "impronta digitale" (o il "DNA") di una distribuzione di probabilità. Ogni distribuzione ne ha una unica, e viceversa. È l'equivalente della Trasformata di Fourier per la probabilità.
- **Perché si usa?** Ha una proprietà magica: la funzione caratteristica della *somma* di variabili indipendenti è semplicemente il *prodotto* delle loro funzioni caratteristiche. (Sommare variabili = Moltiplicare le loro "impronte").
- **Il Teorema di Lévy:** Questo teorema (menzionato nel testo) collega la convergenza delle distribuzioni (quella che ci interessa, $\to_d$) alla convergenza delle loro "impronte digitali" ($\psi_{X_n}(t) \to \psi_X(t)$).

La dimostrazione, quindi, consiste nel:

1. Calcolare la funzione caratteristica della variabile standardizzata (che è una somma).
2. Usare la "proprietà magica" per trasformare la somma in un prodotto, che diventa una potenza $(\dots)^n$.
3. Usare lo sviluppo di Taylor (un'approssimazione) per la funzione all'interno della parentesi.
4. Usare il limite notevole $(1 + a/n)^n \to e^a$ per dimostrare che l'intera espressione converge a $e^{-t^2/2}$.
5. Riconoscere che $e^{-t^2/2}$ è l'"impronta digitale" della Gaussiana Standard $N(0, 1)$.
6. Invocare il Teorema di Lévy per dire: "Se le impronte digitali convergono, allora anche le distribuzioni convergono".

Nell'ambito di questo corso, avremo bisogno di un Teorema del Limite Centrale di portata ben più generale. A tale fine, introduciamo la cosiddetta condizione di Lindeberg.

Condition 65 (Lindeberg). *Sia $\{X_i\}$ una successione di variabili inidpendenti (non necessariamente identicamente distribuite) definite su uno spazio di probabilità $\{\Omega, \mathfrak{S}, \mathbb{P}\}$ e con momento secondo finito; scriviamo $\mu_i = E[X_i]$ e $\sigma_i^2 = Var[X_i]$. Questa sequenza soddisfa la condizione di Lindeberg se per ogni $\epsilon > 0$ si ha*

$$\lim_{n\to\infty} \frac{1}{\sum_{j=1}^n \sigma_j^2} \sum_{i=1}^n E[(X_i - \mu_i)^2 I_{\{|X_i-\mu_i|>\epsilon\sqrt{\sum_{j=1}^n \sigma_j^2}\}}] = 0.$$

(Nota: il testo originale $E[X_i^2 I_{\{X_i^2>\epsilon\sqrt{\sum_{i=1}\sigma_i^2}\}}]$ è stato corretto con la forma standard della condizione, che è più chiara).

Spiegazione Semplice

Perché un altro Teorema? (La condizione di Lindeberg)
Il Teorema 64 è ottimo, ma ha un'ipotesi molto restrittiva: "identicamente distribuite" (i.i.d.). Significa che tutte le X_i devono provenire dalla stessa identica distribuzione (es. tutti lanci dello *stesso* dado).
E se sommiamo cose *diverse*? (es. X_1 è un dado, X_2 è una moneta, X_3 è l'altezza di una persona...).
Il Teorema di Lindeberg (Teorema 66) è un CLT *molto più generale* che funziona anche per variabili **non** identicamente distribuite.
Cosa significa la formula? La formula di Lindeberg è intimidatoria. In parole povere, è una condizione tecnica che assicura che nessuna *singola* variabile X_i sia "troppo importante" e domini la somma.
La somma $\sum(X_i - \mu_i)$ deve essere il risultato di *tanti piccoli* pezzi casuali. Se ci fosse un pezzo (una X_k) la cui varianza σ_k^2 fosse enorme e "schiacciasse" la somma di tutte le altre, il risultato non sarebbe Gaussiano.
La condizione di Lindeberg garantisce che il contributo dei valori "estremi" (le code) di tutte le variabili sia trascurabile rispetto alla varianza totale della somma.

Theorem 66. *Sia $\{X_i\}$ una successione di variabili aleatorie che soddisfa la condizione di Lindeberg. Allora*

$$\frac{\sum_{i=1}^n X_i - \sum_{i=1}^n \mu_i}{\sqrt{\sum_{i=1}^n \sigma_i^2}} \to_d Z \sim N(0,1).$$

Remark 67. *La dimostrazione di questo risultato si basa su approssimazioni successive della funzione caratteristica e non è riportata per brevità. E' però interessante capire il significato della condizione di Lindeberg, che intuitivamente tende ad escludere due casi:*

a) quello in cui la somma delle varianza NON diverga, cioè $\sum_{i=1} \sigma_i^2$ vada ad una costante. In questo caso stiamo stia in pratica sommando solo un numero finito di variabili aleatorie e pertanto il TLC non può valere, a meno che le variabili non siano già in partenza Gaussiane

b) quello in cui la varianza nella coda delle ultime variabili aleatorie sommate sia dello stesso ordine di grandezza di tutte le altre; anche in questo caso il TLC non può valere perché è come se sommassimo due singole variabili, una identificata dalla somma delle prime $n-1$ e l'altra costituita dall'ultima.

Remark 68. *La condizione di Lindeberg è difficile da verificare in pratica e si usa più spesso una meno generale, ma di più facile verifica, la cosiddetta condizione di Lyapunov.*

Condition 69 (Lyapunov)**.** *Sia $\{X_i\}$ una successione di variabili aleatori indipendenti ed a media nulla e definiamo*

$$\tilde{X}_{i,n} = \frac{X_i}{\sqrt{\sum_{j=1}^n E[X_j^2]}}$$

in modo tale che $\sum_{i=1}^n E[\tilde{X}_{i,n}^2] = 1$ per ogni n. La sequenza soddisfa la condizione di Lyapunov se

$$\lim_{n\to\infty} \sum_{i=1}^n E[|\tilde{X}_{i,n}|^{2+\delta}] = 0 \text{ per qualche } \delta > 0.$$

(Nota: il testo originale $\lim_{n\to\infty} max_{i=1,\ldots,n} E[|\tilde{X}_{i,n}|^{2+\delta}] = 0$ è una versione, questa con la somma è un'altra forma comune).

Spiegazione Semplice

Perché ancora un'altra condizione? (Lyapunov)
La condizione di Lindeberg (65) è la condizione teorica perfetta (è quasi "necessaria e sufficiente"), ma è molto difficile da verificare in pratica.
La condizione di Lyapunov è una scorciatoia. È una condizione *più forte* di Lindeberg (cioSè, se vale Lyapunov, vale automaticamente anche Lindeberg), ma è molto *più facile* da calcolare.

In pratica, se riesci a dimostrare che un "momento" leggermente superiore al secondo (come $2+\delta$) va a zero quando normalizzato, hai dimostrato che vale il CLT. È uno strumento pratico per i matematici.

Remark 70. *Anche la condizione di Lyapunov può essere vista come la richiesta che non ci sia nessuna variabile aleatoria con varianza non trascurabile rispetto alla somma di tutte le altre..*

Remark 71. *E' importante discutere la relazione esistente tra teorema del limite centrale e legge dei garndi numeri. Consideriamo il caso di variabili indipendenti X_i, e normalizziamole in modo che abbiano valor medio nullo e varianza 1. Sappiamo che la somma di questa variabili aleatorie ha varianza $Var[\sum_{i=1}^n X_i] = n$, e pertanto $\sum_{i=1}^n X_i = O_p(\sqrt{n})$; in altre parole questa sequenza di somme se divisa per $\sqrt{n}$ vivrà su un compatto con probabilità che può essere resa arbitrariamente vicina ad uno. la legge dei grandi numeri ci dice che dividendo per n abbiamo*

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \to_p 0$$

in un certo senso, il Teorema del Limite Centrale si può pensare come uno zoom sulle fluttuazioni che diventano infinitesimali della media intorno allo zero: riscalando per un fattore $\sqrt{n}$ otteniamo

$$\sqrt{n}\overline{X}_n = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i \to_d N(0,1)$$

. Quindi le fluttuazioni nella somma di n variabili aleatorie indipendenti a valor medio nullo sono di ordine $\sqrt{n}$; le fluttuazioni della media aritmetica sono di ordine $\frac{1}{\sqrt{n}}$ le fluttuazioni nel teorema del limite centrale sono di ordine $O(1)$.

Spiegazione Semplice

Legge dei Grandi Numeri (LLN) vs. Teorema Limite Centrale (CLT)
Questo è un punto cruciale che riassume la differenza tra i due teoremi più importanti:

1. **Legge dei Grandi Numeri (LLN):** Ci dice *dove* la media $\overline{X}_n$ converge.

$$\overline{X}_n \to_p \mu$$

   La media campionaria "collassa" sulla media vera μ. L'errore $(\overline{X}_n - \mu)$ va a zero.

2. **Teorema del Limite Centrale (CLT):** Ci dice *come* la media $\overline{X}_n$ fluttua *mentre* converge. La LLN ci dice che $\overline{X}_n$ si avvicina a μ. L'errore $(\overline{X}_n - \mu)$ è di ordine $O_p(1/\sqrt{n})$, cioè diventa molto piccolo.

   Il CLT fa uno **"zoom"** su questo errore piccolissimo: moltiplica l'errore per $\sqrt{n}$ per "riportarlo in scala".

$$\sqrt{n}(\overline{X}_n - \mu) \to_d N(0,\sigma^2)$$

   Ci dice che questo errore "ingrandito" non va a zero, ma si stabilizza in una perfetta curva Gaussiana.

La LLN ci dà la destinazione, il CLT ci dà la "forma" del percorso e la velocità di convergenza.

Example 72. *(L'uso del Teorema del limite centrale nei sondaggi di opinione.)...*