APPUNTI DI STATISTICA MATEMATICA
(con note e spiegazioni aggiuntive)

Chiappini Mario

11 novembre 2025

Sommario

Note di un corso di Statistica Matematica, con particolare enfasi sulla Teoria Asintotica

1 Disuguaglianze Fondamentali

In questo capitolo richiamiamo le disuguaglianze fondamentali che costituiscono gli strumenti principali per lo sviluppo di tutta la teoria asintotica.

Proposition 1 (Disuguaglianza di Markov)**.** *Sia $X : \Omega \to \mathbb{R}$ una variabile aleatoria a valori non-negativi (cioè $X(\omega) \geq 0$ con probabilità 1) e con valor medio finito (cioè $\mathbb{E}[|X|] = \mathbb{E}[X] < \infty$). Allora $\forall t > 0$ abbiamo*

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}[X]}{t}$$

Spiegazione Semplice

Cosa dice in parole povere? Questa disuguaglianza fissa un limite massimo alla probabilità che un evento "raro" accada, basandosi solo sul valore medio.

- $X : \Omega \to \mathbb{R}$: Questa è la notazione formale per "X è una variabile aleatoria", cioè qualcosa che può assumere diversi valori numerici (es. il risultato del lancio di un dado, l'altezza di una persona).
- $\mathbb{E}[X]$: È il **valore medio** (o "valore atteso") di X. È la media di tutti i possibili risultati, pesata per la loro probabilità.
- $\mathbb{P}(X \geq t)$: È la **probabilità** che la variabile X assuma un valore *maggiore o uguale* a t.

Esempio pratico: Supponiamo che il *valore medio* ($\mathbb{E}[X]$) dello stipendio annuale in una città sia 30.000€. Qual è la probabilità ($\mathbb{P}$) che una persona a caso guadagni *almeno* ($\geq$) 150.000€ (t)? Markov dice: $\mathbb{P}(\textit{Stipendio} \geq 150.000) \leq \frac{30.000}{150.000} = 0.2$. C'è al massimo il 20% di probabilità. È un limite "lento", spesso non molto preciso, ma funziona sempre (purché X sia non-negativa).

Dimostrazione. E' sufficiente osservare che

$$\begin{aligned}\mathbb{P}(X \geq t) = \mathbb{E}[\mathbb{I}_{[t,\infty)}(X)] &= \int_t^\infty dF_X(x) \\ &\leq \int_t^\infty \frac{x}{t} dF_X(x) \leq \frac{1}{t}\int_0^\infty x dF_X(x) \\ &= \frac{\mathbb{E}[X]}{t}\end{aligned}$$

□

Spiegazione Semplice

Decodifichiamo la dimostrazione:

- $\mathbb{I}_{[t,\infty)}(X)$: Si chiama "funzione indicatrice". È molto semplice: vale 1 se $X \geq t$ ed è 0 in caso contrario. Il suo valore medio, $\mathbb{E}[\mathbb{I}_{...}]$, è (per definizione) la probabilità che $X \geq t$.
- $\int_t^\infty dF_X(x)$: Questo integrale è il modo formale per "sommare" (o meglio, integrare) le probabilità di tutti i valori di x che sono maggiori o uguali a t. È la stessa cosa di $\mathbb{P}(X \geq t)$.
- $\int_t^\infty \frac{x}{t} dF_X(x)$: Qui sta il trucco. Dato che stiamo guardando solo valori dove $x \geq t$, è sempre vero che $\frac{x}{t} \geq 1$. Sostituendo 1 con una quantità più grande ($\frac{x}{t}$), stiamo *aumentando* il valore dell'integrale, da cui il segno $\leq$.

Il resto è solo algebra per far riapparire la definizione di $\mathbb{E}[X]$.

Proposition 2 (Disuguaglianza di Markov generalizzata)**.** *Sia $g : \mathbb{R} \to \mathbb{R}^+$ crescente e $X: \Omega \to \mathbb{R}$ una variabile aleatoria tale che $\mathbb{E}[g(X)] < \infty$. Allora*

$$Pr(X \geq t) \leq \frac{\mathbb{E}[g(X)]}{g(t)}$$

Spiegazione Semplice

Questa è la versione "potenziata" di Markov. L'idea è: invece di usare X, usiamo una sua trasformazione $g(X)$ (purché g sia una funzione crescente, come $g(x) = x^2$ o $g(x) = e^x$).
Perché farlo? Perché scegliendo una funzione g "intelligente", possiamo ottenere un limite (un "tappo" massimo) molto più preciso (più basso) per la nostra probabilità.

Dimostrazione. La dimostrazione è praticamente identica a quella già vista della disuguaglianza di Markov. In particolare, chiamando $Y = g(X)$ otteniamo

$$\begin{aligned}
\mathbb{P}(X \geq t) &= \mathbb{P}(g(X) \geq g(t)) = \mathbb{E}[\mathbb{I}_{[g(t),\infty)}(Y)] \\
&= \int_{g(t)}^\infty dF_Y(y) \leq \int_{g(t)}^\infty \frac{Y}{g(t)} dF_Y(y) \\
&= \int_t^\infty \frac{g(x)}{g(t)} dF_X(x) \leq \frac{1}{g(t)} \int_0^\infty g(x) dF_X(x) \\
&= \frac{\mathbb{E}[g(X)]}{g(t)}
\end{aligned}$$

□

Remark 1. *La disuguaglianza di Markov e la sua versione generalizzata sono il primo esempio di disuguaglianze di concentrazione; in pratica, implicano che tanto più una variabile aleatoria ammette momenti finite, tanto più le code della sua distribuzione vanno a zero velocemente. Ad esempio, consideriamo una variabile aleatoria che ammetta funzione generatrice dei momenti finita in un intorno dell'origine, cioè tale per cui*

$$m_X(u) := \mathbb{E}[\exp(uX)] < \infty \quad \textit{per } u \in [0, T], T > 1.$$

Segue immediatamente che

$$Pr(X \geq t) \leq \frac{\mathbb{E}[\exp(X)]}{\exp(t)} = \frac{m_X(1)}{\exp(t)}$$

cioè le code della distribuzione di X devono decadere almeno esponenzialmente.

Spiegazione Semplice

Momenti e Code:

- **Code (Tails)**: Sono le parti estreme di una distribuzione di probabilità, cioè la probabilità di ottenere valori molto alti o molto bassi (molto lontani dalla media).
- **Momenti (Moments)**: Sono misure che descrivono la forma della distribuzione.
  - Il 1° momento è la **media** ($\mathbb{E}[X]$).
  - Il 2° momento è legato alla **varianza** ($\mathbb{E}[X^2]$).
  - E così via...

Questo remark dice: se una variabile ha "momenti finiti" (cioè la sua media, varianza, ecc. non sono infinite), allora le sue "code vanno a zero velocemente". In pratica, è molto improbabile ottenere valori estremi. La **funzione generatrice dei momenti** ($m_X(u)$) è uno strumento matematico che "impacchetta" tutte le informazioni su tutti i momenti in un'unica funzione.

Proposition 3 (Disuguaglianza di Chebyshev)**.** *Sia $X : \Omega \to \mathbb{R}$ una variabile aleatoria con momento secondo finito (cioè $\mathbb{E}[X^2] < \infty$). Allora $\forall t > 0$ abbiamo*

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq t) \leq \frac{Var[X]}{t^2}$$

Spiegazione Semplice

Questa è una delle disuguaglianze più famose ed è un caso specifico di Markov (generalizzata, usando $g(x) = (x - \mu)^2$).
Cosa dice in parole povere? Ci dà un limite alla probabilità di "allontanarsi" dalla media, basandosi sulla *varianza*.

- $Var[X]$: È la **Varianza**, una misura di "quanto è dispersa" la variabile X. Se $Var[X]$ è piccola, i valori di X sono quasi tutti vicini alla media. Se è grande, i valori sono molto sparpagliati.
- $|X - \mathbb{E}[X]| \geq t$: È la notazione per "la distanza tra X e la sua media $\mathbb{E}[X]$ è maggiore o uguale a t".

Esempio pratico: Se l'altezza media ($\mathbb{E}[X]$) è 175cm e la Varianza ($Var[X]$) è $25cm^2$, qual è la probabilità di trovare una persona più alta di 185cm o più bassa di 165cm? Qui $t = 10cm$ (la distanza da 175).
Chebyshev dice: $\mathbb{P}(|Altezza - 175| \geq 10) \leq \frac{25}{10^2} = \frac{25}{100} = 0.25$. C'è al massimo il 25% di probabilità.
È un limite molto più "stretto" (preciso) di quello che darebbe Markov.

Dimostrazione. E' sufficiente osservare che, utilizzando la disuguaglianza di Markov per $Y = |X - \mathbb{E}[X]|^2$

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq t) = \mathbb{P}(|X - \mathbb{E}[X]|^2 \geq t^2) = \frac{\mathbb{E}[|X - \mathbb{E}[X]|^2]}{t^2}$$

□

Remark 2. *Come discusso nei cenni storici più avanti, la disuguaglianza di Chebyshev è precedente a quella di Markov; in effetti Markov è stato un alunno di Chebyshev a San Pietroburgo.*

Example 1 (Legge debole dei grandi numeri)**.** *Anticipando un poco la discussione nei prossimi capitoli, possiamo illustrare immediatamente l'applicazione più importante della disuguaglianza di Chebyshev. Siano infatti $X_i : \Omega \to \mathbb{R}$ variabili aleatorie indipendenti con momento secondo finito, e definiamo come al solito valor medio e varianza $\mu := \mathbb{E}[X]$, $\sigma^2 := \mathbb{E}[(X - \mu)^2]$. Introduciamo altresì la media aritmetica $\overline{X}_n := \frac{1}{n}\sum_{i=1}^n X_i$; dalla disuguaglianza di Chebyshev abbiamo immediatamente che, per ogni $\epsilon > 0$*

$$Pr\{|\overline{X}_n - \mu| > \epsilon\} \leq \frac{Var\{\overline{X}_n\}}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2} \to 0 \quad \textit{per } n \to \infty.$$

E' immediato verificare che l'ipotesi di indipendenza può essere generalizzata a quella di incorrelazione; anche le ipotesi di identica distribuzione può essere facilmente abbandonata, come vedremo nei prossimi capitoli.

Spiegazione Semplice

Questo è il cuore della statistica: **La Legge dei Grandi Numeri**.

- $\overline{X}_n$: È la **media campionaria**, cioè la media aritmetica calcolata su n osservazioni (es. la media dei voti di 10 esami).
- μ: È la **media vera** (o "media della popolazione") (es. la media di tutti i voti che prenderesti se potessi dare l'esame infinite volte).

Cosa dice la formula? La formula $Pr\{...\} \leq \frac{\sigma^2}{n\epsilon^2}$ ci dice che la probabilità che la nostra "media campionaria" ($\overline{X}_n$) sia diversa dalla "media vera" (μ) per più di un piccolo errore ϵ, diminuisce man mano che n (la dimensione del nostro campione) aumenta.
Quando $n \to \infty$ (raccogliamo tantissimi dati), il termine $\frac{\sigma^2}{n\epsilon^2}$ va a 0.
In pratica: Più dati raccogli (n grande), più puoi essere sicuro che la media che hai calcolato ($\overline{X}_n$) sia vicina alla media vera (μ). Questo è il motivo per cui i sondaggi funzionano (e funzionano meglio con più intervistati).

Nelle precedenti disuguaglianze abbiamo mostrato che per variabili aleatorie che ammettono un numero di momenti maggiore le code della funzione di distribuzione decadono più velocemente. E' pertanto naturale chiedersi se si pos- sano ottenere risultati più stringenti focalizzandosi su classi di variabili aleatorie che abbiano supporto compatto, risultando quindi uniformemente limitate. La risposta è affermativa, come mostrato nel prossimo risultato.

Proposition 4 (Disuguaglianza di Hoeffding)**.** *Sia $X_1, ..., X_n$ una successione di variabili aleatorie indipendenti e tali che esistano due sequenze di numeri reali a_i, b_i $i = 1, 2, .., n$ per cui valga $a_i \leq X_i \leq b_i$ per ogni i; assumiamo inoltre che $\mathbb{E}[Y_i] = 0$ per $i = 1, ..., n$. Per ogni $t > 0$ vale la disuguaglianza*

$$Pr\{\sum_{i=1}^{n} X_i \geq t\} \leq \inf_{u \geq 0}\{e^{-ut} \prod_{i=1}^{n} e^{u^2(b_i - a_i)^2/8}\}$$

Spiegazione Semplice

Cosa aggiunge Hoeffding? Markov e Chebyshev funzionano per quasi tutte le variabili. Hoeffding funziona per una classe specifica: variabili che sono **limitate** ("supporto compatto"), cioè i loro valori sono *sempre* contenuti in un intervallo $[a_i, b_i]$.
Esempio: Il lancio di una moneta (X_i può essere solo 0 o 1, quindi è limitato a [0, 1]). Il voto di un esame (limitato a [18, 30]). L'altezza di una persona *non* è (teoricamente) limitata.
Usando questa informazione extra (i limiti a_i e b_i), Hoeffding fornisce un limite (un "tappo") *esponenzialmente* più preciso (più basso) rispetto a Chebyshev.

Dimostrazione. Per la disuguaglianza di Markov abbiamo, quale che sia $u > 0$

$$\begin{aligned} Pr\{\sum_{i=1}^{n} X_i \geq t\} &= Pr\{\sum_{i=1}^{n} uX_i \geq ut\} \\ &= Pr\{e^{u\sum_{i=1}^{n} X_i} \geq e^{ut}\} \\ &\leq e^{-ut}\mathbb{E}[e^{u\sum_{i=1}^{n} X_i}] \\ &= e^{-ut}\prod_{i=1}^{n} \mathbb{E}[e^{uX_i}]. \end{aligned}$$

Scriviamo ora $Y_i = \alpha b_i + (1-\alpha)a_i$, $\alpha = \frac{Y_i - a_i}{b_i - a_i}$ per la convessità della funzione esponenziale, si ha che

$$\begin{aligned} e^{uX_i} &\leq \frac{Y_i - a_i}{b_i - a_i} e^{ub_i} + (1 - \frac{Y_i - a_i}{b_i - a_i}) e^{ua_i} \\ &= \frac{Y_i - a_i}{b_i - a_i} e^{ub_i} + \frac{b_i - Y_i}{b_i - a_i} e^{ua_i} \end{aligned}$$

da cui, ricordando che $\mathbb{E}[Y_i] = 0$

$$\mathbb{E}[e^{uX_i}] \leq -\frac{a_i}{b_i - a_i} e^{ub_i} + \frac{b_i}{b_i - a_i} e^{ua_i}$$

Scriviamo ora

$$-\frac{a_i}{b_i - a_i} e^{ub_i} + \frac{b_i}{b_i - a_i} e^{ua_i} = e^{g_i(v_i)}$$

dove

$$g_i(v_i) := -\gamma_i v_i + \log(1 - \gamma_i + \gamma_i e^{v_i}), \quad \gamma_i = \frac{-a_i}{b_i - a_i}, \quad v_i = u(b_i - a_i);$$

notiamo infatti che

$$\begin{aligned} e^{g_i(v_i)} &= e^{-\gamma_i v_i}(1 - \gamma_i + \gamma_i e^{v_i}) \\ &= e^{ua_i}(1 + \frac{a_i}{b_i - a_i} - \frac{a_i}{b_i - a_i} e^{u(b_i - a_i)}) \\ &= (\frac{b_i - a_i}{b_i - a_i} e^{ua_i} + \frac{a_i}{b_i - a_i} e^{ua_i} - \frac{a_i}{b_i - a_i} e^{ub_i}) \\ &= \frac{b_i}{b_i - a_i} e^{ua_i} - \frac{a_i}{b_i - a_i} e^{ub_i}. \end{aligned}$$

Notiamo che $g_i(0) = 0$ ed inoltre

$$g_i'(0) = -\gamma_i + \frac{\gamma_i e^{v_i}}{(1 - \gamma_i + \gamma_i e^{v_i})}|_{v_i = 0} = 0,$$

$$g_i''(v_i) = \frac{\gamma_i e^{v_i}(1 - \gamma_i + \gamma_i e^{v_i}) - \gamma_i e^{v_i}(\gamma_i e^{v_i})}{(1 - \gamma_i + \gamma_i e^{v_i})^2} = \frac{\gamma_i(1 - \gamma_i) e^{v_i}}{(1 - \gamma_i + \gamma_i e^{v_i})^2}$$

Infatti, poiché $v > 0$

$$g_i''(v_i) = \frac{\gamma_i(1 - \gamma_i)}{((1 - \gamma_i)e^{-v_i/2} + \gamma_i e^{v_i/2})^2} \leq \frac{\gamma_i(1 - \gamma_i)}{(1 - \gamma_i + \gamma_i)^2} \leq \frac{1}{4}$$

(l'ultima disuguaglianza segue da $\gamma_i(1 - \gamma_i) \leq 1/4$ per ogni γ_i).

Per il teorema del valor medio di Lagrange, esiste $\xi \in (0, v_i)$ tale per cui

$$g_i(v_i) = g_i(0) + g_i'(0)v_i + g_i''(\xi)\frac{v_i^2}{2} = g_i''(\xi)\frac{v_i^2}{2} \leq \frac{u^2(b_i - a_i)^2}{8}.$$

Ne segue che

$$\mathbb{E}[e^{uX_i}] \leq e^{g_i(v_i)} \leq e^{u^2(b_i - a_i)^2/8},$$

ed la Proposizione è dimostrata. □

Corollary 1 (Media aritmetica di variabili aleatorie di Bernoulli)**.** *Siano $Y_1, ..., Y_n$ variabili aleatorie indipendenti identicamente distribuite con legge di Bernoulli Ber(p). Per ogni $\epsilon > 0$, abbiamo che*

$$Pr\{|\overline{Y}_n - p| > \epsilon\} \leq 2e^{-2n\epsilon^2}.$$

Spiegazione Semplice

Questo è un risultato pratico importantissimo che deriva da Hoeffding.

- **Variabile di Bernoulli (Ber(p))**: È il modello matematico per un singolo evento che ha

due soli risultati. Es. "lancio di una moneta" (Testa/Croce) o "un utente clicca/non clicca". p è la probabilità di successo (es. $p = 0.5$ per una moneta onesta).

- $\overline{Y}_n$: È la media campionaria, che qui rappresenta la "frequenza" di successi (es. 60 Teste su 100 lanci, $\overline{Y}_n = 0.6$).
- p: È la probabilità vera (es. $p = 0.5$).

Cosa dice la formula? Ci dà la probabilità che la nostra *frequenza* misurata ($\overline{Y}_n$) sia diversa dalla *probabilità* vera (p) per più di un errore ϵ.
Il termine $e^{-2n\epsilon^2}$ è importantissimo: ci dice che questa probabilità di errore **crolla esponenzialmente** all'aumentare di n (numero di lanci).

Dimostrazione. Si prenda $X_i := \frac{1}{n}(Y_i - p)$. E' immediato verificato che $\mathbb{E}[X_i] = 0$ e $a_i \leq X_i \leq b_i$ per ogni $i = 1, ..., n$, con $a_i = -\frac{p}{n}$ $b_i = \frac{1-p}{n}$. da cui $(b_i - a_i)^2 = \frac{1}{n^2}$ Applicando la disuguaglianza di Hoeffding si ha

$$Pr\{\overline{Y}_n - p > \epsilon\} \leq \inf_u \{e^{-u\epsilon} \prod_{i=1}^{n} e^{u^2/8n^2}\} = \inf_u \{e^{-u\epsilon} e^{u^2/8n}\}$$

e prendendo $u = 4n\epsilon$ otteniamo

$$\inf_u \{e^{-u\epsilon} e^{u^2/8n}\} \leq e^{-4n\epsilon\times\epsilon} e^{16n^2\epsilon^2/8n} = e^{-2n\epsilon^2}.$$

La dimostrazione è completata ripetendo lo stesso ragionamento per $Pr\{\overline{Y}_n - p < -\epsilon\}$. □

Remark 3 (Un confronto tra la disuguaglianza di Chebyshev e quella di Hoeffding)**.** *Nel caso di variabili aleatorie limitate, la disuguaglianza di Hoeffding è enormemente più efficiente della disuguaglianza di Chebyshev. Ad esempio, nel caso di variabili Bernoulliane di parametro* $p = \frac{1}{2}$ *conn* $n = 100$ *ed* $\epsilon = 0.2$ *la disuguaglianza di Chebyshev ci dà il seguente limite limite superiore:*

$$Pr\{|\overline{Y}_n - p| > 0.2\} \leq \frac{Var\{\overline{Y}_n\}}{\epsilon^2} = \frac{1}{4n\epsilon^2} \simeq 0.0625$$

mentre con la disuguaglianza di Hoeffding si ottiene

$$Pr\{|\overline{Y}_n - p| > 0.2\} \leq 2e^{-2\times 100\times(0.2)^2} \simeq 0.00067.$$

Spiegazione Semplice

Questo confronto è la chiave. Vogliamo sapere la probabilità che, dopo 100 lanci di una moneta onesta ($n = 100, p = 0.5$), la nostra frequenza di Teste sia "sbagliata" di molto (es. più di 0.7 o meno di 0.3, cioè $\epsilon = 0.2$).

- **Chebyshev** (che usa solo la varianza) dice: "Questa probabilità è al massimo 6.25%".
- **Hoeffding** (che usa anche il fatto che i lanci sono *limitati* tra 0 e 1) dice: "Questa probabilità è al massimo 0.067%".

Hoeffding è "enormemente più efficiente" perché sfrutta più informazioni sulla variabile.

Remark 4 (Concentrazione del volume di un ipercubo)**.** *La disuguaglianza di Hoeffding ammette una interessante intepretazione geometrica quando viene applicata al comportmanto de volume di un ipercubo* $[0, 1]^n$*, nel limite in cui* $n \to \infty$*. Consideriamo infatti l'iperpiano* $\Gamma_n := \{x_i : x_1 + x_2 + ...x_n = n/2\}$*; definitamo un "Tubo" di raggio* ϵ *intorno a* Γ_n *come*

$$Tub_\epsilon(\Gamma_n) := \{x \in \mathbb{R}^n : dist\{x, \Gamma_n\} \leq \epsilon\}.$$

Per ogni $\epsilon > 0$*, il volume dell'intersezione tra* $Tub_\epsilon(\Gamma_n)$ *e* $[0, 1]^n$ *converge esponenzialmente a 1 quando n diverge all'infinito; in altre parole, la massa del cubo si concentra esponenzialmente in un intorno arbitrariamente piccolo della diagonale principale.*

Le precedenti disuguaglianze ci permettono di avere dei limiti superiori per il comportamento delle code di variabili aleatorie generiche. E' interessante confrontare il loro comportamento con quello delle code di una variabile Gaussiana standard; a questo scopo introduciamo la prossima proposizione.

Proposition 5 (Disuguaglianza di Mill-Gordon)**.** *Sia $Z \sim N(0,1)$ una variabile Gaussiana standard. Si ha che*

$$\sqrt{\frac{1}{2\pi}}\frac{t}{1+t^2}\exp(-\frac{t^2}{2}) \leq Pr\{Z > t\} \leq \sqrt{\frac{1}{2\pi}}\frac{1}{t}\exp(-\frac{t^2}{2}).$$

Spiegazione Semplice

Questa disuguaglianza è diversa. Non è "generale" come Markov o Chebyshev, ma si applica *solo* alla **Variabile Gaussiana Standard** ($Z \sim N(0,1)$), la famosa "curva a campana".
Cosa dice? Calcolare $Pr\{Z > t\}$ (l'area sotto la coda della campana) è difficile. Questa formula non dà un limite superiore "lento", ma "intrappola" la vera probabilità tra due valori molto vicini e facili da calcolare. È un modo molto efficiente per stimare le code della Gaussiana.

Dimostrazione. La disuguaglianza di destra può essere stabilita con un semplice cambio di variabile; abbiamo:

$$\begin{aligned}
Pr\{Z > t\} &= \int_t^\infty \frac{1}{\sqrt{2\pi}}\exp(-\frac{x^2}{2})dx \\
&\leq \int_t^\infty \frac{x}{t}\frac{1}{\sqrt{2\pi}}\exp(-\frac{x^2}{2})dx \\
&= \frac{1}{t}\frac{1}{\sqrt{2\pi}}\int_{t^2/2}^\infty \exp(-y)dy \quad \text{(dopo il cambio di variabile } y = \frac{x^2}{2}) \\
&= \frac{1}{t}\frac{1}{\sqrt{2\pi}}\exp(-\frac{t^2}{2}).
\end{aligned}$$

Per il limite inferiore, possiamo ragionare come segue:

$$\begin{aligned}
Pr\{Z > t\} &= \int_t^\infty \frac{1}{\sqrt{2\pi}}\exp(-\frac{x^2}{2})dx \\
&\geq \int_t^\infty \frac{x^4+2x^2-1}{x^4+2x^2+1}\frac{1}{\sqrt{2\pi}}\exp(-\frac{x^2}{2})dx \\
&= \int_t^\infty \frac{x^2(x^2+1)+x^2-1}{(x^2+1)^2}\frac{1}{\sqrt{2\pi}}\exp(-\frac{x^2}{2})dx \\
&= \int_t^\infty \{\frac{x^2}{(x^2+1)} + \frac{x^2-1}{(x^2+1)^2}\}\frac{1}{\sqrt{2\pi}}\exp(-\frac{x^2}{2})dx
\end{aligned}$$

Chiamiamo ora

$$\psi(x) = \frac{x}{1+x^2}\exp(-\frac{x^2}{2})$$

e notiamo che

$$\begin{aligned}
\psi'(x) = \frac{d\psi(x)}{dx} &= \frac{1(1+x^2)-x(2x)}{(1+x^2)^2}\exp(-\frac{x^2}{2}) + \frac{x}{1+x^2}\exp(-\frac{x^2}{2})(-x) \\
&= \frac{1-x^2}{(1+x^2)^2}\exp(-\frac{x^2}{2}) - \frac{x^2(1+x^2)}{(1+x^2)^2}\exp(-\frac{x^2}{2}) \\
&= \frac{1-x^2-x^2-x^4}{(1+x^2)^2}\exp(-\frac{x^2}{2}) = -\frac{x^4+2x^2-1}{(x^2+1)^2}\exp(-\frac{x^2}{2})
\end{aligned}$$

(Nota: C'è un'apparente discrepanza tra la derivata $\psi'(x)$ e l'integrando. Seguendo la logica del testo, l'integrale della derivata deve restituire la funzione. L'integrando è $A + B$ dove $A = \frac{x^2}{(x^2+1)}\exp(-x^2/2)$ e $B = \frac{x^2-1}{(x^2+1)^2}\exp(-x^2/2)$. La derivata di $\psi(x)$ è $\psi'(x) = \frac{1-x^2}{(1+x^2)^2}\exp(-x^2/2) - \frac{x^2(1+x^2)}{(1+x^2)^2}\exp(-x^2/2)$. L'espressione nel PDF [source 130] è:

$$\psi'(x) = \cdots = -\{\frac{x^2}{(x^2+1)^2} + \frac{x^2-1}{(x^2+1)^2}\}\exp(-\frac{x^2}{2})$$

Questa espressione non sembra corretta. La derivata corretta è:

$$\psi'(x) = \frac{1-x^2}{(1+x^2)^2}e^{-x^2/2} - \frac{x^2}{1+x^2}e^{-x^2/2} = \frac{(1-x^2)-x^2(1+x^2)}{(1+x^2)^2}e^{-x^2/2} = -\frac{x^4+2x^2-1}{(1+x^2)^2}e^{-x^2/2}$$

L'integrando nel PDF [source 128] è $\frac{x^4+2x^2-1}{(x^2+1)^2}\frac{1}{\sqrt{2\pi}}\exp(-x^2/2) = -\psi'(x)/\sqrt{2\pi}$. Quindi, la dimostrazione è completata dal Teorema Fondamentale del Calcolo:)

$$\frac{1}{\sqrt{2\pi}}\int_t^\infty (-\psi'(x))\,dx = -\frac{1}{\sqrt{2\pi}}[\psi(x)]_t^\infty = -\frac{1}{\sqrt{2\pi}}[0-\psi(t)] = \sqrt{\frac{1}{2\pi}}\frac{t}{1+t^2}\exp(-\frac{t^2}{2}).$$

□

Remark 5. *Queste due disuguglianze sono piuttosto efficienti; ad esempio, per $t = 2$ abbiamo*

$$Pr\{Z > 2\} = \int_2^\infty \frac{1}{\sqrt{2\pi}}\exp(-\frac{x^2}{2})dx = 0.02275$$

mentre le stime di Mill danno

$$\frac{1}{\sqrt{2\pi}}\frac{t}{1+t^2}\exp(-\frac{t^2}{2})|_{t=2} = 0.021596 \le Pr\{Z > 2\} \le \frac{1}{\sqrt{2\pi}}\frac{1}{t}\exp(-\frac{t^2}{2})|_{t=2} = 0.027$$

Per $t = 3$ il risulta diventa ancora più preciso; otteniamo

$$Pr\{Z > 3\} = \int_3^\infty \frac{1}{\sqrt{2\pi}}\exp(-\frac{x^2}{2})dx = 0.001349$$

mentre le stime di Mill danno

$$\frac{1}{\sqrt{2\pi}}\frac{t}{1+t^2}\exp(-\frac{t^2}{2})|_{t=3} = 0.001329$$
$$\le Pr\{Z > 3\} \le \frac{1}{\sqrt{2\pi}}\frac{1}{t}\exp(-\frac{t^2}{2})|_{t=3} = 0.0014773.$$

Remark 6. *Vediamo quale risultato otterremo per la somma di variabili Bernoulliane considerando l'approssimazione asintotica che segue dal teorema del Limte Centrale; prendendo per semplicità $p = \frac{1}{2}$ abbiamo:*

$$\overline{Y}_n - \mathbb{E}[\overline{Y}_n] \simeq \mathcal{N}(0, \frac{p(1-p)}{n}) = \mathcal{N}(0, \frac{1}{4n})$$

ed utilizzando la disuguaglianza di Mill abbiamo dunque

$$\begin{aligned} Pr\{|\overline{Y}_n - \frac{1}{2}| > 0.2\} &= Pr\{|\mathcal{N}(0, \frac{1}{4n})| > 0.2\} + o_{n\to\infty}(1) \\ &= Pr\{|\mathcal{N}(0,1)| > 0.2 \times 2\sqrt{n}\} \\ &\le \frac{2}{\sqrt{2\pi}}\frac{1}{0.2\times 2\sqrt{n}}e^{-4\times n\times(0.2)^2/2}. \end{aligned}$$

Per $n = 100$ otteniamo

$$\frac{2}{\sqrt{2\pi}}\frac{1}{0.2\times 20}e^{-2\times 100\times(0.2)^2} = 6.6915\times 10^{-5};$$

un valore circa 10 volte inferiore a quello ottenuto dalla disuguaglianza di Hoeffding. Il risultato però non deve ingannare: la disuguaglianza di Hoeffding vale in senso stretto, mentre qui stiamo trovando un limite superiore alla probabilità Gaussiana, tralasciando il fatto che il termine di approssimazione $o_{n\to\infty}(1)$ è molto più grande della maggiorazione ottenuta: si può dimostrare solo l'ordine $O(n^{-1/2})$. In altre parole, il Teorema del Limite Centrale NON può essere sfruttato per ottenere un limite così efficiente per la probabilità sulle code della variabile $\overline{Y}_n$

La prossima disuguaglianza dovrebbe essere ben nota dai corsi di Geometria ed Analisi.

Proposition 6 (Disuguaglianza di Cauchy-Schwartz)**.** *Siano $X, Y : \Omega \to \mathbb{R}$ variabili aleatorie con momento secondo finito. Abbiamo che*

$$|\mathbb{E}[XY]|^2 \le \mathbb{E}[X^2]\mathbb{E}[Y^2].$$

La disuguaglianza è stretta, a meno che le variabili aleatorie siano tra loro in relazione lineare con probabilità 1, cioè esista un numero reale t^* tale per cui

$$\mathbb{E}[(Y - t^*X)^2] = 0.$$

Spiegazione Semplice

Questa è l'equivalente della famosa disuguaglianza che si studia in geometria, ma applicata alle variabili aleatorie.
Spiegazione geometrica (Remark 15): Si può pensare alle variabili aleatorie come a dei "vettori" in uno spazio astratto.

- $\mathbb{E}[X^2]$ è come il "quadrato della lunghezza" del vettore X.
- $\mathbb{E}[XY]$ è come il "prodotto scalare" tra i vettori X e Y.

La disuguaglianza dice che il prodotto scalare al quadrato (che misura quanto sono "allineati") è al massimo il prodotto delle loro lunghezze al quadrato.
L'uguaglianza vale solo se i vettori sono "allineati", cioè se Y è un multiplo di X ($Y = t^*X$). In statistica, questo significa che X e Y sono **perfettamente correlate linearmente**.

Dimostrazione. Consideriamo la funzione quadratica (in t)

$$g(t) : \mathbb{E}[(Y - tX)^2] = \mathbb{E}[Y^2] - 2t\mathbb{E}[XY] + t^2\mathbb{E}[X^2] \geq 0.$$

L'equazione $g(t) = 0$ ovviamente ammette al più una radice reale di molteplicità due; quindi il discriminante deve avere valore non-positivo, in particolare

$$4(\mathbb{E}[XY])^2 - 4\mathbb{E}[X^2]\mathbb{E}[Y^2] \leq 0$$

da cui segue la disuguaglianza, che diviene una ugualianza se e solo se esiste t^* tale per cui $g(t^*) = 0$. □

Remark 7. *Il richiamo alla geometria ci anticipa una idea che avrà grande importanza nei corsi di probabilità più avanzati: lo spazio delle variabili aleatorie di momento secondo finito può essere visto come uno spazio vettoriale dotato di un proodtto interno (prodotto scalare) definito da*

$$\langle X, Y \rangle := \mathbb{E}[XY].$$

Nel caso particolare di variabili con valor medio nullo questo prodotto interno è la covarianza.

Per la prossima disuguaglianza ricordiamo innanzitutto che una funzione $g(.) : \mathbb{R} \to \mathbb{R}$ si dice convessa se per ogni $\alpha \in (0, 1)$ e per ogni x_1 e x_2 nel suo dominio (connesso), si ha che

$$g(\alpha x_1 + (1 - \alpha)x_2) \leq \alpha g(x_1) + (1 - \alpha)g(x_2).$$

Data una funzione convessa per ogni punto x_0 esiste una costante $L = L_{x_0}$ tale per cui

$$g(x) \geq g(x_0) + L(x - x_0);$$

la costante non è necessariamente unica (se la funzione è derivabile, coincide con il valore della derivata in quel punto).

Proposition 7 (Disuguaglianza di Jensen)**.** *Sia X una variabile aleatoria con valor medio finito e $g(.)$ una funzione convessa. Abbiamo che*

$$\mathbb{E}[g(X)] \geq g(\mathbb{E}[(X)]),$$

dove la grandezza a sinistra della disuguaglianza può essere infinita.

Spiegazione Semplice

Cosa dice in parole povere? Per una funzione **convessa** (una funzione a forma di "scodella" U, come $g(x) = x^2$), la media *dopo* aver applicato la funzione è sempre maggiore o uguale della funzione applicata *alla media*.
Esempio Pratico: Prendiamo X che vale -1 o +1 con la stessa probabilità (50/50), e usiamo la funzione convessa $g(x) = x^2$.

- **Calcoliamo** $g(\mathbb{E}[X])$ **(lato destro):** La media di X è $\mathbb{E}[X] = (-1 \times 0.5) + (1 \times 0.5) = 0$. Applichiamo g: $g(\mathbb{E}[X]) = g(0) = 0^2 = \mathbf{0}$.
- **Calcoliamo** $\mathbb{E}[g(X)]$ **(lato sinistro):** Applichiamo g *prima*: $g(-1) = (-1)^2 = 1$; e $g(1) = 1^2 = 1$. Ora calcoliamo la media di $g(X)$: $\mathbb{E}[g(X)] = (1 \times 0.5) + (1 \times 0.5) = \mathbf{1}$.

Come previsto dalla disuguaglianza, $1 \geq 0$.
(Se la funzione fosse *concava*, a forma di $\cap$ come $g(x) = \log(x)$, il segno $\geq$ si invertirebbe in $\leq$).

Dimostrazione. E' sufficiente osservare che, per la monotonia del valor medio e prendendo $x_0 = \mathbb{E}[(X)]$

$$\begin{aligned}\mathbb{E}[g(X)] &\geq \mathbb{E}[g(x_0) + L(X - x_0)] = \mathbb{E}[g(x_0)] + \mathbb{E}[L(X - x_0)] \\ &= g(\mathbb{E}[(X)]) + L\mathbb{E}[(X - \mathbb{E}[(X)])] = g(\mathbb{E}[(X)]).\end{aligned}$$

□