Azzolini Riccardo 2020-05-05

Legge dei grandi numeri

1 Convergenza delle proporzioni empiriche

Sia X una variabile aleatoria che può assumere m valori, indicati con $1, \dots, m$ (per semplicità, qui si considera il caso discreto), rispettivamente con probabilità $p_1, \dots, p_m$. Sia $\{X_n\}$ una successione di variabili aleatorie indipendenti e distribuite identicamente a X (ovvero un campionamento della popolazione rappresentata da X).

Scegliendo un valore $k \in \{1, \dots, m\}$ tra i possibili assunti dalle variabili, si definisce per ogni X_i un'altra variabile aleatoria, Y_i, che assume valore 1 se e solo se $X_i = k$, e 0 altrimenti, cioè che indica il verificarsi dell'evento $\{X_i = k\}$:

$$Y_i = \begin{cases} 1 & \text{se } X_i = k \\ 0 & \text{altrimenti} \end{cases} \qquad i = 1, 2, \dots$$

Considerando poi un campione di ampiezza n, composto dalle prime n variabili $X_1, \dots,$ X_n della successione $\{X_n\}$, la media campionaria delle corrispondenti $Y_1, \dots, Y_n$

$$\bar{p}_k^{(n)} = \frac{1}{n} \sum_{i=1}^{n} Y_i$$

conta il numero di occorrenze dell'evento osservate nelle prove, in rapporto al numero n di prove effettuate, e prende il nome di **proporzione empirica**. Essendo un rapporto tra casi favorevoli e casi possibili, ottenuto da variabili X_i distribuite ugualmente a X, la proporzione empirica $\bar{p}_k^{(n)}$ può essere considerata una stima della probabilità $P\{X = k\} = p_k$.

Si è già visto che, in generale, la media campionaria (qui $\bar{p}_k^{(n)}$) è anch'essa una variabile aleatoria, il cui valore medio coincide con la media dell'intera popolazione: in questo caso, se si definisce sulla popolazione X una variabile Y analoga alle Y_i,

$$Y = \begin{cases} 1 & \text{se } X = k \\ 0 & \text{altrimenti} \end{cases}$$

essa ha valore medio

$$E(Y) = 1 \cdot P\{X = k\} + 0 \cdot P\{X \neq k\} = P\{X = k\} = p_k$$

e ciò conferma che $\bar{p}_k^{(n)}$ è, appunto, una stima della probabilità p_k.

Tuttavia, anche se $E(\bar{p}_k^{(n)}) = E(Y) = p_k$, i singoli valori di $\bar{p}_k^{(n)}$ variano casualmente (intorno alla media p_k) al variare del campione. Intuitivamente, ci si aspetterebbe che, all'aumentare di n, queste variazioni tendano a diventare sempre più piccole, fino ad avere, al limite per $n \to +\infty$, che $\bar{p}_k^{(n)} = p_k$: si dice che la variabile aleatoria $\bar{p}_k^{(n)}$ **converge** a p_k (per $n \to +\infty$). Si pone allora il problema di dimostrare tale convergenza.

Osservazione: È interessante notare che qui si ha una variabile aleatoria ($\bar{p}_k^{(n)}$) che converge a un valore deterministico, non casuale: la probabilità p_k, che è una proprietà intrinseca del fenomeno modellato dalla variabile aleatoria X. Ad esempio, se si considera il lancio di una moneta, la proporzione empirica di teste ottenute in n lanci varia casualmente, ma, per n grandi, tende a $\frac{1}{2}$, un valore fissato dalle caratteristiche proprie della moneta.

2 Convergenza di variabili aleatorie

Prima di poter dimostrare la convergenza delle proporzioni empiriche, bisogna dare una definizione formale della convergenza di una successione di variabili aleatorie $\{X_n\}$ a un'altra variabile aleatoria X (da non confondere con le X_n e la X del problema precedente, nel quale la convergenza è considerata quella delle $\{\bar{p}_k^{(n)}\}$ a p_k).

Esistono varie definizioni, più o meno forti:

- Se $F_n(t)$ e $F(t)$ sono le funzioni di ripartizione di X_n e X, rispettivamente, si potrebbe chiedere che
$$\lim_{n\to+\infty} F_n(t) = F(t)$$
cioè, ad esempio, per X_n e X continue con densità $f_n(s)$ e $f(s)$:
$$\lim_{n\to+\infty} \int_{-\infty}^{t} f_n(s)\, ds = \int_{-\infty}^{t} f(s)\, ds$$
Questa è chiamata *convergenza in distribuzione*, ed è la stessa considerata nel teorema del limite centrale.
- Si potrebbe chiedere la *convergenza dei momenti* di ordine r:
$$\lim_{n\to+\infty} E(|X_n - X|^r) = 0$$
- Qui si scelgono invece altre due nozioni di convergenza: la **convergenza quasi certa** e la **convergenza in probabilità**, presentate in seguito.

2.1 Convergenza quasi certa

Definizione: Siano $\{X_n\}$ una successione di variabili aleatorie $(X_1, X_2, \ldots)$, e X una variabile aleatoria. Si dice che $\{X_n\}$ **converge** a X **quasi certamente**, $X_n \xrightarrow{\text{q.c.}} X$, se e solo se l'insieme degli $\omega \in \Omega$ tali che

$$\lim_{n \to +\infty} X_n(\omega) = X(\omega)$$

ha probabilità 1:

$$P\Big\{\omega \;\Big|\; \lim_{n \to +\infty} X_n(\omega) = X(\omega)\Big\} = 1$$

Ciò significa che

$$\lim_{n \to +\infty} X_n(\omega) = X(\omega)$$

vale per ogni $\omega \in \Omega$, a meno di un insieme di misura nulla.

2.2 Convergenza in probabilità

Definizione: Siano $\{X_n\}$ una successione di variabili aleatorie, e X una variabile aleatoria. Si dice che $\{X_n\}$ **converge** a X **in probabilità**, $X_n \xrightarrow{P} X$, se e solo se, per ogni numero $\eta > 0$ fissato, si ha

$$\lim_{n \to +\infty} P\{|X_n - X| > \eta\} = 0$$

In altre parole, $X_n \xrightarrow{P} X$ se e solo se, per η arbitrariamente piccoli e n sufficientemente grandi, la probabilità che X_n si discosti da X per più di η tende a 0.

Si può dimostrare che la convergenza quasi certa implica la convergenza in probabilità,

$$X_n \xrightarrow{\text{q.c.}} X \implies X_n \xrightarrow{P} X$$

ma non viceversa. Quindi, la convergenza quasi certa è una condizione più forte rispetto a quella in probabilità.

3 Disuguaglianza di Chebyshev

Un altro elemento necessario per dimostrare la convergenza delle proporzioni empiriche è la **disuguaglianza di Chebyshev**: per ogni $\eta > 0$,

$$P\{|X - E(X)| > \eta\} \leq \frac{\mathrm{Var}(X)}{\eta^2}$$

Dimostrazione: Per dimostrare quest'uguaglianza, si ricorre all'artificio di definire la variabile aleatoria

$$Y = \begin{cases} \eta^2 & \text{se } |X - E(X)| > \eta \\ 0 & \text{altrimenti} \end{cases}$$

e si osserva che $(X - E(X))^2 \geq Y$, perché:

- se $|X - E(X)| > \eta$, allora $Y = \eta^2$, e

$$(X - E(X))^2 > \eta^2 = Y$$

- altrimenti, $Y = 0$, e dunque

$$(X - E(X))^2 \geq 0 = Y$$

Per la monotonia del valore medio,[1] la disuguaglianza $(X - E(X))^2 \geq Y$ può essere applicata al calcolo della varianza di X:

$$\begin{aligned}
\mathrm{Var}(X) &= E\big((X - E(X))^2\big) && \text{(definizione di varianza)} \\
&\geq E(Y) && \text{(monotonia del valore medio)} \\
&= \eta^2 P\{|X - E(X)| > \eta\} + 0P\{|X - E(X)| \leq \eta\} \\
&= \eta^2 P\{|X - E(X)| > \eta\}
\end{aligned}$$

Da qui, è sufficiente portare η^2 al denominatore per ottenere la disuguaglianza di Chebyshev,

$$\begin{aligned}
\mathrm{Var}(X) &\geq \eta^2 P\{|X - E(X)| > \eta\} \\
\frac{\mathrm{Var}(X)}{\eta^2} &\geq P\{|X - E(X)| > \eta\}
\end{aligned}$$

completando così la dimostrazione.

3.1 Altre forme

L'evento $\{|X - E(X)| > \eta\}$ equivale a

$$\begin{aligned}
\{X - E(X) < -\eta\} \cup \{X - E(X) > \eta\} &= \{X < E(X) - \eta\} \cup \{X > E(X) + \eta\} \\
&= \{X \notin (E(X) - \eta,\ E(X) + \eta)\}
\end{aligned}$$

quindi la disuguaglianza di Chebyshev afferma che la probabilità che X assuma un valore fuori dall'intervallo $(E(X) - \eta,\ E(X) + \eta)$ è minore o uguale a $\frac{\mathrm{Var}(X)}{\eta^2}$. Di conseguenza, più è piccola la varianza, minore è la probabilità che X assuma valori fuori da tale

[1] La proprietà di monotonia del valore medio afferma che, in generale, se $P\{X \geq Y\} = 1$, allora $E(X) \geq E(Y)$.

intervallo (e infatti, intuitivamente, una varianza più piccola corrisponde a valori più concentrati vicino alla media).

La disuguaglianza di Chebyshev può anche essere espressa nella forma relativa all'evento complementare:

$$P\{|X - E(X)| \le \eta\} \ge 1 - \frac{\operatorname{Var}(X)}{\eta^2}$$

3.2 Precisione della stima

La caratteristica più importante della disuguaglianza di Chebyshev è che si applica a qualsiasi distribuzione di probabilità di cui siano noti il valore medio e la varianza.

D'altro canto, proprio per questo, la stima della probabilità $P\{|X - E(X)| > \eta\}$ che essa fornisce tende a essere una maggiorazione piuttosto grossolana, poco precisa. Perciò, quando possibile, conviene usare approssimazioni migliori, come ad esempio quella data dal teorema del limite centrale.

4 Legge dei grandi numeri

Il teorema che formalizza e generalizza la convergenza delle proporzioni empiriche è la **legge dei grandi numeri**:

Teorema: Sia $\{X_n\}$ una successione di variabili aleatorie indipendenti e aventi tutte la stessa legge, caratterizzata da un valore medio μ e da una varianza finita σ^2. Allora, posto

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n}(X_1 + \cdots + X_n)$$

(la media campionaria delle prime n variabili della successione), si ha che $\overline{X}_n \xrightarrow{\text{q.c.}} \mu$ (*legge forte dei grandi numeri*[2]), e quindi anche $\overline{X}_n \xrightarrow{P} \mu$ (*legge debole dei grandi numeri*).

4.1 Dimostrazione della legge debole

Per semplicità, si dà solo la dimostrazione della legge debole, $\overline{X}_n \xrightarrow{P} \mu$, che è il caso più facile.

[2] La legge forte dei grandi numeri presenta molte versioni. Qui è riportata quella di base, mentre la formulazione più generale è addirittura oggetto di ricerca.

Per prima cosa, si calcola $E(\overline{X}_n)$, sfruttando la linearità del valore medio e ricordando che, per ipotesi, $E(X_1) = \cdots = E(X_n) = \mu$:

$$
\begin{aligned}
E(\overline{X}_n) &= E\left(\frac{1}{n}(X_1 + \cdots + X_n)\right) \\
&= \frac{1}{n} E(X_1 + \cdots + X_n) \\
&= \frac{1}{n}\big(E(X_1) + \cdots + E(X_n)\big) \\
&= \frac{1}{n}(\underbrace{\mu + \cdots + \mu}_{n \text{ volte}}) = \frac{1}{n} n\mu = \mu
\end{aligned}
$$

Il calcolo di $\operatorname{Var}(\overline{X}_n)$ avviene in modo analogo, poiché, per ipotesi, le X_i sono indipendenti (quindi vale la linearità della varianza) e hanno $\operatorname{Var}(X_1) = \cdots = \operatorname{Var}(X_n) = \sigma^2$:

$$
\begin{aligned}
\operatorname{Var}(\overline{X}_n) &= \operatorname{Var}\left(\frac{1}{n}(X_1 + \cdots + X_n)\right) \\
&= \frac{1}{n^2} \operatorname{Var}(X_1 + \cdots + X_n) \\
&= \frac{1}{n^2}\big(\operatorname{Var}(X_1) + \cdots + \operatorname{Var}(X_n)\big) \\
&= \frac{1}{n^2}(\underbrace{\sigma^2 + \cdots + \sigma^2}_{n \text{ volte}}) = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}
\end{aligned}
$$

Adesso, si applica la disuguaglianza di Chebyshev:

$$
\begin{gathered}
P\{|\overline{X}_n - E(\overline{X}_n)| > \eta\} \le \frac{\operatorname{Var}(\overline{X}_n)}{\eta^2} \\
P\{|\overline{X}_n - \mu| > \eta\} \le \frac{\sigma^2}{n\eta^2}
\end{gathered}
$$

Infine, poiché

$$
\lim_{n \to +\infty} \frac{\sigma^2}{n\eta^2} = 0
$$

deve essere

$$
\lim_{n \to +\infty} P\{|\overline{X}_n - \mu| > \eta\} = 0
$$

cioè $X_n \xrightarrow{P} \mu$.

5 Esempio: proporzioni empiriche per una moneta

Si suppone di lanciare una moneta, non sapendo se essa sia equilibrata o meno. La probabilità p di ottenere testa in un lancio non è quindi nota; se la moneta fosse equilibrata, sarebbe $p = \frac{1}{2}$.

Per studiare la probabilità p, si effettuano n lanci, i cui risultati sono rappresentati dalle variabili aleatorie $X_1, \dots, X_n$, indipendenti e ugualmente distribuite: potendo assumere solo i valori 1 (testa) e 0 (croce), esse sono variabili di Bernoulli $B(1, p)$, ciascuna avente media $E(X_i) = p$.

La somma $X_1 + \dots + X_n$ conta il numero di teste ottenute negli n lanci, quindi la media campionaria

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

è la proporzione empirica che stima la probabilità p. Si osserva che in questo caso non è stato necessario definire separatamente le variabili

$$Y_i = \begin{cases} 1 & \text{se } X_i = k \\ 0 & \text{altrimenti} \end{cases} \qquad i = 1, \dots, n$$

perché le X_i assumono già solo i valori 1 (testa, l'evento che si vuole contare) e 0 (l'evento che invece non deve essere contato, cioè croce).

Applicando la legge dei grandi numeri, si avrebbe che $\overline{X}_n \longrightarrow E(X_i) = p$ per $n \to +\infty$, ma, in pratica, si può fare solo un numero n finito di lanci. Allora, non sarà possibile determinare il valore esatto di p, ma solo stimarlo con $\overline{X}_n$, e quantificare l'errore $|\overline{X}_n - p|$ che si commette impiegando tale stima.

Un modo per quantificare l'errore è fornito dalla disuguaglianza di Chebyshev:

$$P\{|\overline{X}_n - E(\overline{X}_n)| > \eta\} \le \frac{\operatorname{Var}(\overline{X}_n)}{\eta^2}$$

Infatti:

- Siccome $\overline{X}_n$ è la media campionaria, $E(\overline{X}_n) = E(X_i) = p$.
- Essendo $X_1 + \dots + X_n$, in quanto somma di Bernoulli $B(1, p)$ indipendenti, una variabile binomiale $B(n, p)$, si ha

$$\operatorname{Var}(X_1 + \dots + X_n) = np(1 - p)$$

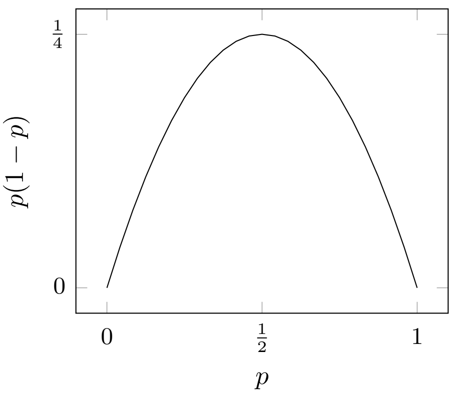
e quindi

$$\begin{aligned}\mathrm{Var}(\overline{X}_n) &= \mathrm{Var}\left(\frac{1}{n}(X_1 + \cdots + X_n)\right) \\ &= \frac{1}{n^2}\,\mathrm{Var}(X_1 + \cdots + X_n) \\ &= \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}\end{aligned}$$

Inserendo questi risultati nella disuguaglianza di Chebyshev, si ottiene:

$$P\{|\overline{X}_n - p| > \eta\} \leq \frac{p(1-p)}{n\eta^2}$$

Il valore a destra dell'uguaglianza non può essere calcolato, perché dipende da p, che è sconosciuta (è proprio il valore che si vuole stimare), ma, studiando il grafico della parabola $p(1-p) = -p^2 + p$,



si nota che è possibile introdurre la maggiorazione $p(1-p) \leq \frac{1}{4}$:

$$P\{|\overline{X}_n - p| > \eta\} \leq \frac{p(1-p)}{n\eta^2} \leq \frac{1}{4n\eta^2}$$

Così, adesso, si può determinare, ad esempio, un numero n di lanci (ampiezza del campione) per il quale risulti minore o uguale a 0.5 la probabilità che la media campionaria $\overline{X}_n$ si discosti da p per più di $\eta = 0.1$:

$$\begin{aligned}P\{|\overline{X}_n - p| > 0.1\} \leq \frac{1}{4n \cdot 0.1^2} &\leq 0.5 \\ \frac{1}{0.04n} &\leq 0.5 \\ 0.04n &\geq \frac{1}{0.5} \\ 0.04n &\geq 2 \\ n &\geq \frac{2}{0.04} = \frac{1}{0.02} = 50\end{aligned}$$

In altre parole, questo risultato significa che, dopo 50 lanci, sarà garantita una probabilità ≤ 50 % che la "vera" probabilità p di ottenere testa si discosti dalla stima fornita dalla proporzione empirica per più di 0.1.

6 Problema: uso della disuguaglianza di Chebyshev

Problema: Una variabile aleatoria X ha valore medio $\mu = 3$ e varianza $\sigma^2 = 2$. Mediante la disuguaglianza di Chebyshev, determinare una maggiorazione per le seguenti probabilità:

1. $P\{|X - 3| \geq 2\}$
2. $P\{|X - 3| \geq 1\}$
3. $P\{|X - 3| \leq 1.5\}$

Soluzioni:

1. Siccome la probabilità da stimare è della forma $P\{|X - E(X)| \geq \eta\}$, con $E(X) = \mu = 3$ e $\eta = 2$, si applica direttamente la disuguaglianza di Chebyshev:

$$P\{|X - 3| \geq 2\} \leq \frac{\sigma^2}{\eta^2} = \frac{2}{2^2} = \frac{1}{2} = 0.5$$

2. Il calcolo è analogo al precedente, ma con $\eta = 1$:

$$P\{|X - 3| \geq 1\} \leq \frac{\sigma^2}{\eta^2} = \frac{2}{1^2} = 2$$

   Si osserva che la stima ottenuta in questo caso è assolutamente inutile, perché una probabilità è sempre ≤ 1, e quindi anche ≤ 2.

3. Per applicare la disuguaglianza di Chebyshev a quest'ultima probabilità, bisogna passare all'evento complementare:

$$P\{|X - 3| > 1.5\} \leq \frac{2}{1.5^2} = \frac{2}{\left(\frac{3}{2}\right)^2} = \frac{2 \cdot 2^2}{3^2} = \frac{8}{9}$$

   Quindi:

$$P\{|X - 3| \leq 1.5\} = 1 - P\{|X - 3| > 1.5\} \geq 1 - \frac{8}{9} = \frac{1}{9} \approx 0.111$$

7 Problema: automobili prodotte

Problema: Il numero di automobili prodotte da una fabbrica in una settimana è una variabile aleatoria X con valore medio $\mu = 500$ e varianza $\sigma^2 = 100$. Stimare la probabilità che, questa settimana, la produzione sia compresa tra 400 e 600 automobili.

Soluzione: Osservando che

$$400 = 500 - 100 = \mu - 100 \qquad 600 = 500 + 100 = \mu + 100$$

la probabilità richiesta può essere espressa come

$$\begin{aligned} P\{400 \leq X \leq 600\} &= P\{\mu - 100 \leq X \leq \mu + 100\} \\ &= P\{-100 \leq X - \mu \leq 100\} \\ &= P\{|X - \mu| \leq 100\} \end{aligned}$$

quindi si può applicare la forma della disuguaglianza di Chebyshev relativa all'evento complementare:

$$\begin{aligned} P\{|X - \mu| \leq \eta\} &\geq 1 - \frac{\sigma^2}{\eta^2} \\ P\{|X - 500| \leq 100\} &\geq 1 - \frac{100}{100^2} = 1 - \frac{1}{100} = \frac{99}{100} = 0.99 \end{aligned}$$

8 Problema: clienti di un concessionario

Problema: Il numero di clienti che visitano un concessionario di auto il sabato mattina è una variabile aleatoria X con valore medio $\mu = 18$ e deviazione standard $\sigma = 2.5$. Con quale probabilità si può asserire che il numero di clienti è compreso tra 8 e 28?

Soluzione: Il calcolo è del tutto analogo al problema precedente; bisogna solo fare attenzione al fatto che il testo del problema indica la deviazione standard σ, mentre nella disuguaglianza di Chebyshev va usata la varianza σ^2.

$$\begin{aligned} P\{8 \leq X \leq 28\} &= P\{18 - 10 \leq X \leq 18 + 10\} \\ &= P\{|X - \underbrace{18}_{\mu}| \leq 10\} \\ &\geq 1 - \frac{2.5^2}{10^2} = 1 - \left(\frac{1}{4}\right)^2 \\ &= 1 - \frac{1}{16} = \frac{15}{16} = 0.9375 \end{aligned}$$

9 Problema: uso della disuguaglianza di Chebyshev

Problema: Una variabile aleatoria X ha valore medio $\mu = 6$ e deviazione standard $\sigma = \sqrt{2}$; trovare una stima della probabilità che X assuma valori compresi tra 4.5 e 7.5.

Soluzione: Il calcolo è ancora lo stesso:

$$\begin{aligned} P\{4.5 \leq X \leq 7.5\} &= P\{6 - 1.5 \leq X \leq 6 + 1.5\} \\ &= P\{|X - 6| \leq 1.5\} \\ &\geq 1 - \frac{2}{1.5^2} = 1 - \frac{2 \cdot 2^2}{3^2} \\ &= 1 - \frac{8}{9} = \frac{1}{9} \approx 0.111 \end{aligned}$$

10 Problema: confronto tra probabilità esatta e stima

Problema: Una variabile aleatoria X ha la densità di probabilità

$$f(x) = \begin{cases} 2e^{-2x} & \text{se } x \geq 0 \\ 0 & \text{se } x < 0 \end{cases}$$

Sapendo che il valore medio e la varianza valgono rispettivamente $\mu = \frac{1}{2}$ e $\sigma^2 = \frac{1}{4}$,

1. calcolare $P\{|X - \mu| \geq 1\}$
2. trovare una stima per $P\{|X - \mu| \geq 1\}$ con la disuguaglianza di Chebyshev

e confrontare i due risultati.

Soluzioni:

1. Si riconosce che X è una variabile aleatoria esponenziale di parametro $\lambda = 2$. Allora, si può usare la corrispondente funzione di ripartizione[3]

$$F_X(x) = \begin{cases} 1 - e^{-2x} & \text{se } x \geq 0 \\ 0 & \text{se } x < 0 \end{cases}$$

[3] In alternativa, come al solito, si potrebbe integrare direttamente la densità sull'intervallo considerato.

per calcolare la probabilità richiesta:

$$\begin{aligned}
P\{|X-\mu| \geq 1\} &= 1 - P\{|X-\mu| < 1\} \\
&= 1 - P\{\mu - 1 < X < \mu + 1\} \\
&= 1 - P\left\{-\frac{1}{2} < X < \frac{3}{2}\right\} \\
&= 1 - F_X\left(\frac{3}{2}\right) + F_X\left(-\frac{1}{2}\right) \\
&= 1 - 1 + e^{-2\cdot\frac{3}{2}} + 0 \\
&= e^{-3} \approx 0.04979
\end{aligned}$$

2. Applicando la disuguaglianza di Chebyshev, si trova

$$P\{|X-\mu| \geq 1\} \leq \frac{\frac{1}{4}}{1^2} = \frac{1}{4} = 0.25$$

che è una stima molto grossolana del valore esatto, 0.04979.