7 Il TLC

A questo punto del percorso, disponiamo di vari strumenti relativi alla v.a. normale. Essi ci consentiranno di precisare la questione di approssimazione introdotta nel paragrafo 1.3. In tale sezione abbiamo visto che la *distribuzione binomiale* si può approssimare con una opportuna *distribuzione normale.* Ma

> *in che senso* la distribuzione binomiale *è approssimata* dalla distribuzione normale ?

Precisare tale approssimazione è un requisito indispensabile per effettuare concretamente l'approssimazione. In particolare relativamente al problema dei sondaggi che ci siamo posti all'inizio e che costituisce il filo conduttore.

7.1 Esplorazioni: i parametri -attività-

La nostra investigazione sull'approssimazione della binomiale inizia dall'esame dei valori da attribuire ai parametri della variabile aleatoria normale.
Precisamente, consideriamo una v.a. binomiale di parametri n e p fissati, dove n rappresenta il numero di prove e p la probabilità di successo nella prova.

Per quali valori dei parametri μ e σ^2 della v.a. normale, la densità approssima la distribuzione binomiale?

La forma dei due grafici suggerisce che le medie (normale e binomiale) debbano coincidere. Ossia ci aspettiamo valga

$$\boxed{\mu = np}$$

Resta dunque da individuare un valore per il parametro σ^2. È ciò che ci proponiamo di fare con l'attività seguente.

Nota. Per semplicità diciamo di confrontare il grafico della distribuzione binomiale con quello della densità normale. Ma distribuzione di probabilità e densità sono oggetti matematici diversi. Preciseremo più avanti in che senso vada considerato tale confronto.

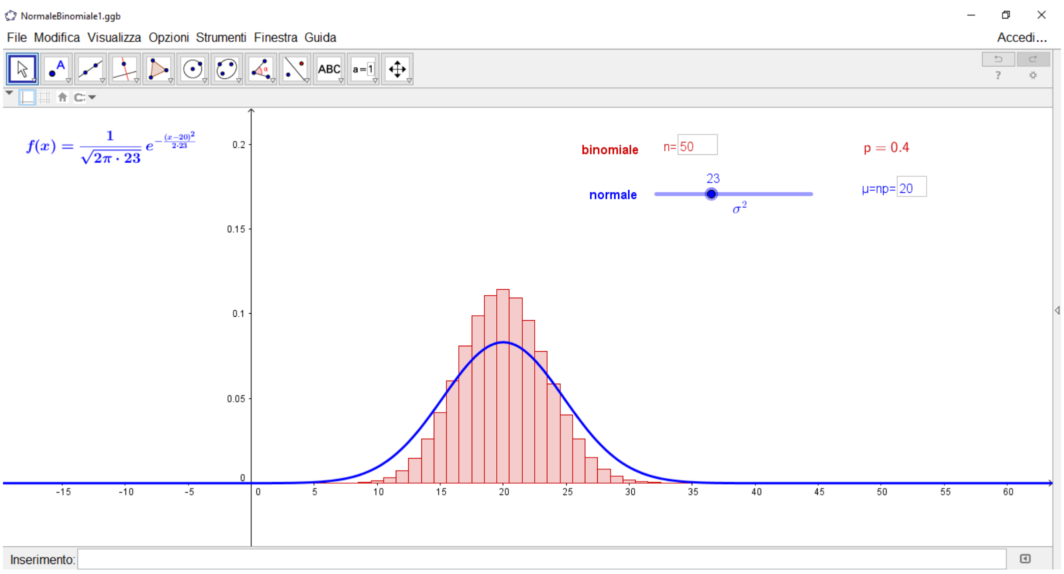
Utilizza il file Geogebra NormaleBinomiale1.ggb per rispondere alle sequenti questioni.

Considera le v.a. binomiali di media fissata $np = 20$ e per i seguenti valori di n: $n = 25$, $n = 50$, $n = 100$ e $n = 200$.
Considera poi la v.a. normale che ha media $\mu = np = 20$.

a) *In ciascuno dei casi indicati, determina un valore di σ^2 per il quale la curva normale sembra approssimare la distribuzione binomiale.*

b) *Quale relazione ti sembra possa esserci tra il valore trovato di σ^2 e quello dei parametri n, p della binomiale?*

Come si utilizza il file NormaleBinomiale1.ggb



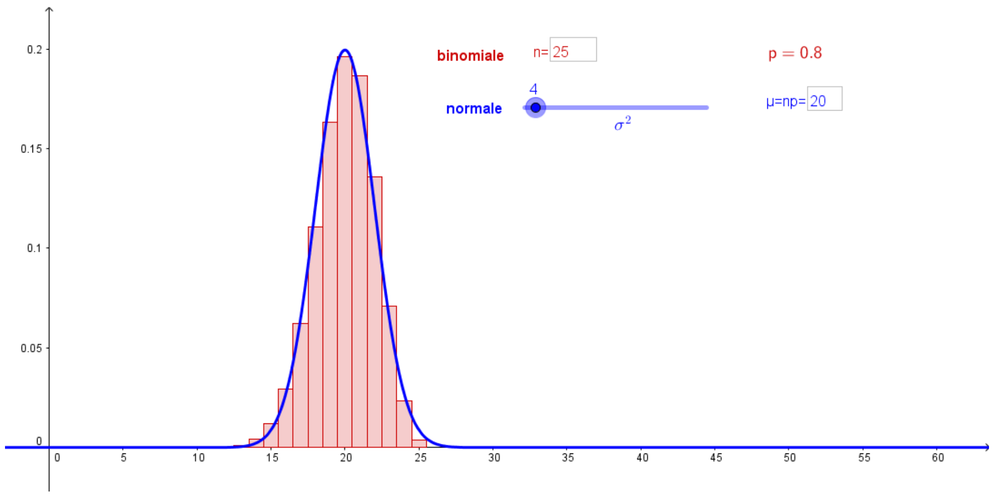
Il file dispone di:

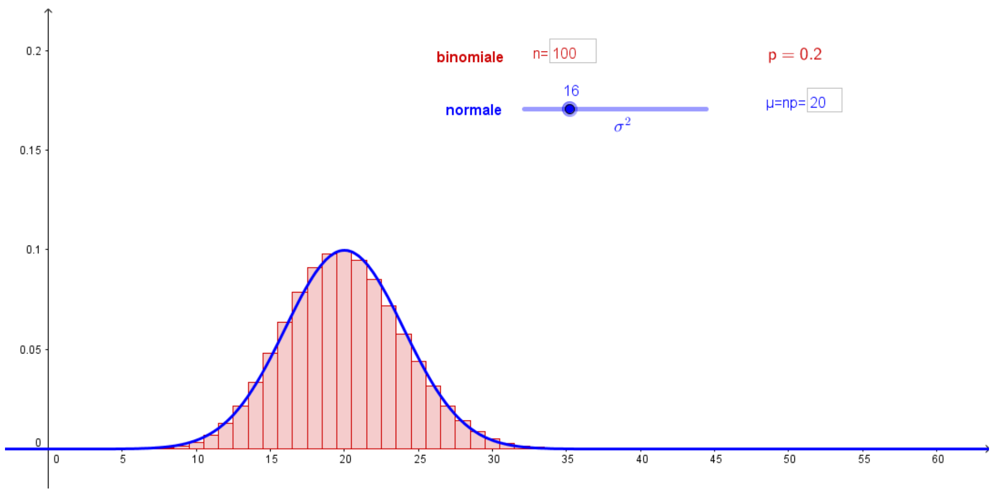
- una casella di inserimento per il parametro n della binomiale: a seconda del valore di n qui inserito, ottieni (automaticamente) il grafico della distribuzione binomiale;
- uno slider per il parametro σ^2 della normale: al variare di σ^2 (muovi il cursore) si ottengono i corrispondenti grafici della densità normale.

Risoluzione

a) Esaminando i grafici, si dovrebbe osservare che si ha una "buona" approssimazione intorno a:

- $\sigma^2 = 4$ quando $n = 25$
- $\sigma^2 = 12$ quando $n = 50$
- $\sigma^2 = 16$ quando $n = 100$
- $\sigma^2 = 18$ quando $n = 200$





b) Facendo un po' di conti, pare che il valore di σ^2, individuato nel punto a) nei vari casi, sia proprio uguale a $np(1-p)$. Ma quest'ultima espressione rappresenta la varianza della v.a. binomiale. Pertanto, almeno nei casi esaminati, sembra che si abbia una "buona" approssimazione quando le varianze delle due v.a. coincidono.

Conclusioni

Il risultato congetturato nell'attività precedente vale più in generale.

> Se la varianza della v.a. normale è uguale a quella della v.a. binomiale (e le medie delle due v.a. coincidono), si ha una "buona" approssimazione della distribuzione binomiale. Ossia per
> $$\sigma^2 = np(1-p)$$

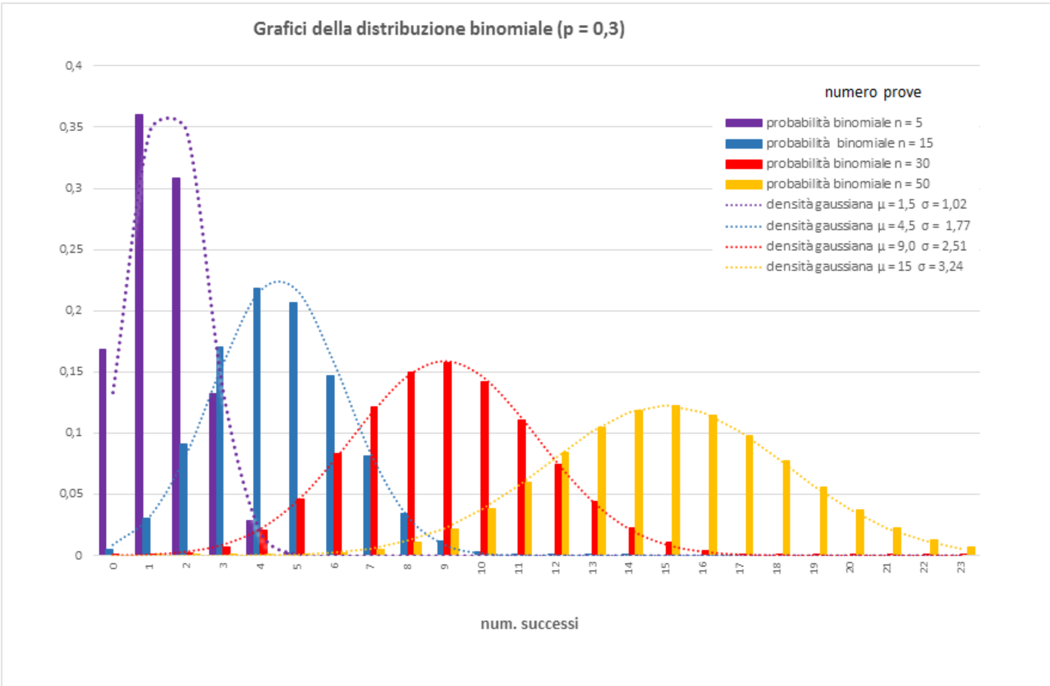
Attenzione però: come vedremo nella prossima attività, l'uguaglianza dei parametri della binomiale e della normale *non è sufficiente* per avere una "buona" approssimazione.

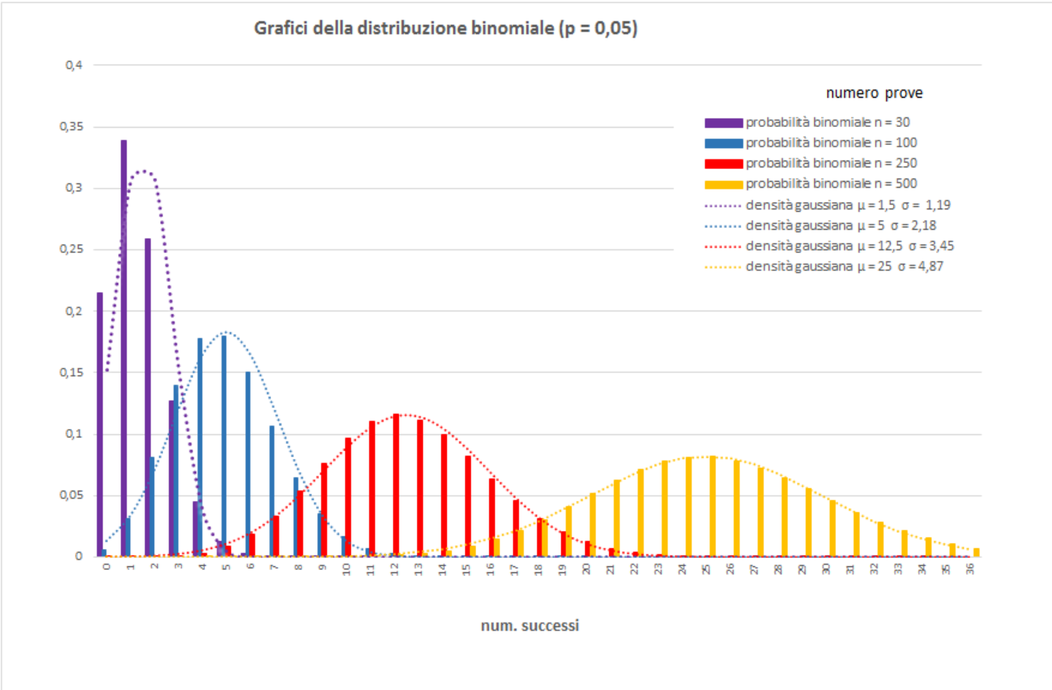
7.2 Esplorazioni: il numero di prove -attività-

Una volta stabilito il legame tra i parametri μ e σ^2 della v.a. normale e i parametri n e p della v.a. binomiale, sondiamo come varia l'approssimazione al variare del numero di prove.

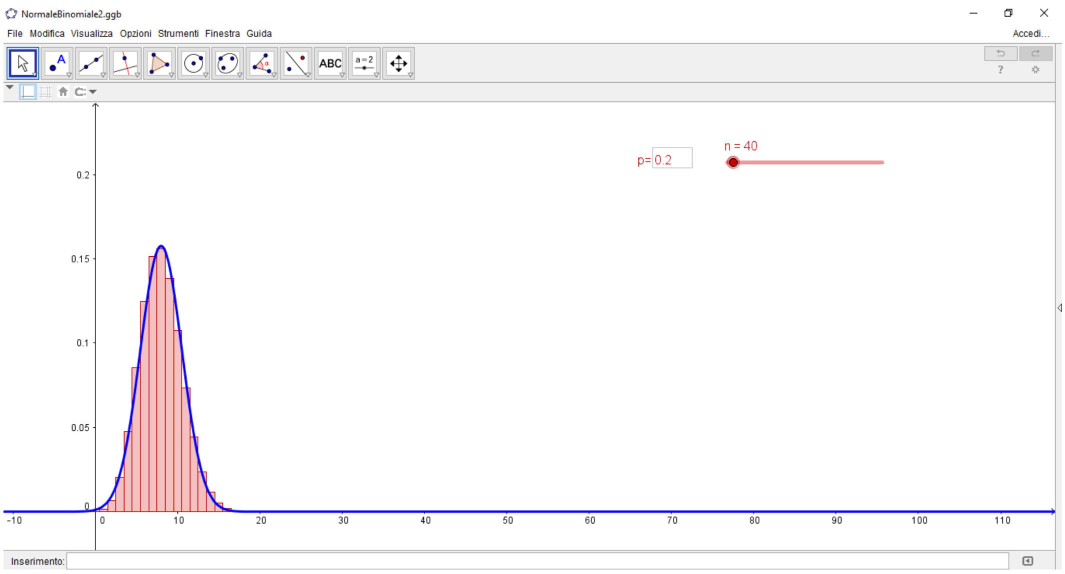
Nelle seguenti figure fissiamo prima $p = 0,3$ e poi $p = 0,05$. In entrambi i casi confrontiamo i grafici (realizzati con Excel) relativi alla v.a. binomiale e alla v.a. normale, per valori crescenti di n.

Osserviamo che con $p = 0,05$ occorre un n più grande, rispetto al caso $p = 0,3$, perché la distribuzione binomiale assomigli ad una curva "a campana".





Usa il file Geogebra NormaleBinomiale2.ggb per effettuare alcune prove.



I grafici esaminati suggeriscono che l'approssimazione è "buona" per n "grande".

Conclusioni

Ciò vale in generale, cioè

per n "*grande*", è "buona" l'approssimazione della distribuzione binomiale con la normale (che ha la stessa media e varianza della binomiale).

7.3 La sostanza del teorema

Le due conclusioni a cui siamo giunti negli ultimi due paragrafi si possono sintetizzare nel seguente teorema.

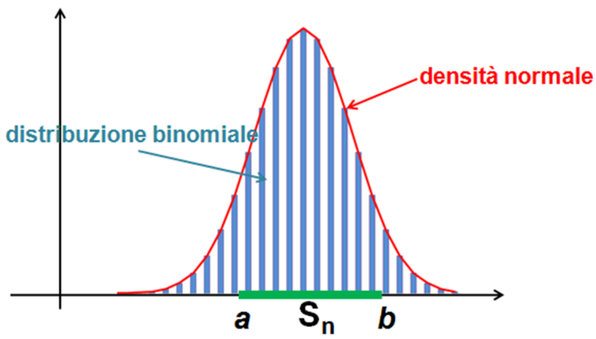
TEOREMA LIMITE CENTRALE -TLC (formulazione semplificata)

Sia S_n la v.a. binomiale di parametri n e p, dove n rappresenta il numero di prove e p la probabilità di successo nella prova.
Sia X la v.a. normale che ha media μ e varianza σ^2 uguali rispettivamente alla media e alla varianza di S_n. Ossia

$$\mu = np \quad \text{e} \quad \sigma^2 = np(1-p)$$

Allora per n "grande", vale l'approssimazione:

$$\mathbf{P}(\mathbf{a} \leq \mathbf{S_n} \leq \mathbf{b}) \simeq \mathbf{P}(\mathbf{a} \leq \mathbf{X} \leq \mathbf{b})$$



Nota. In figura stiamo confrontando una *distribuzione di probabilità* (quella binomiale) con una *densità di probabilità* (quella normale). La figura suggerisce che i loro valori siano "vicini", ma esse sono oggetti matematici di tipo diverso, mentre nell'enunciato del TLC intervengono solo valori di probabilità. Pertanto, a rigore, dovremmo interpretare la distribuzione binomiale come una funzione costante a tratti sull'intervallo reale $[a; b]$: con questo accorgimento le due probabilità confrontate nel TLC si possono leggere in figura come aree dei corrispondenti sottografici in $[a; b]$.

7.4 Applicazioni: finalmente i sondaggi

Torniamo al problema guida proposto nel paragrafo 1.1.

Una popolazione costituita da **10.000** individui è chiamata a votare tra due candidati, diciamo A e B. Mediante un sondaggio effettuato su un campione della popolazione, si è stimato che la probabilità che l'**individuo** sia favorevole ad A è del **40**%.

Sulla base di ciò, si vuole stimare il numero **F** di individui che voteranno A. Precisamente, quanto vale

$$\mathbf{P}(\mathbf{3900} \leq \mathbf{F} \leq \mathbf{4100})\ ?$$

Ora mediante il TLC possiamo risolverlo in modo computazionalmente meno articolato:

1. modellizziamo la situazione mediante la distribuzione binomiale, come già fatto all'inizio del percorso; essa ha parametri $n = 10.000$ e $p = 0,4$
2. per il TLC, il numero di individui F si approssima mediante la v.a. normale X di parametri
$$\mu = np = 4.000 \quad \text{e} \quad \sigma^2 = np(1-p) = 2400$$
3. calcoliamo quindi la probabilità richiesta

$$\mathbf{P}(\mathbf{3900} \leq \mathbf{F} \leq \mathbf{4100}) \underset{\substack{\uparrow \\ \textit{per il TLC}}}{\simeq} P(3900 \leq X \leq 4100) \underset{\substack{\uparrow \\ \textit{standardizzando } X}}{=} P\left(\frac{3900-4000}{\sqrt{2400}} \leq Z \leq \frac{4100-4000}{\sqrt{2400}}\right)$$

$$\simeq P(-2,04 \leq Z \leq 2,04) \underset{\substack{\uparrow \\ \textit{ricorrendo alle tavole}}}{\simeq} 0,9586$$

Pertanto[30] la probabilità che la percentuale di cittadini favorevoli ad A sia compresa tra il 39% e il 41% dell'intera popolazione è circa $\boxed{0,9586}$.

[30] Nell'ultimo passaggio, conviene calcolare $P(-2,04 \leq Z \leq 2,04)$ sfruttando la simmetria della densità normale e dell'intervallo di variabilità di Z. Pertanto tale probabilità è uguale a $2P(0 \leq Z \leq 2,04) = 2(P(Z \leq 2,04) - P(Z \leq 0))$. Dalle tavole si deduce che $P(Z \leq 2,04) \simeq 0,9793$; invece per $P(Z \leq 0)$ **non serve** ricorrere alle tavole.

Osservazione: i due approcci a confronto

- Il TLC permette di risolvere il problema iniziale senza dover svolgere molti conti. Invece il ricorso al modello binomiale, come si accorse De Moivre, comporta parecchie valutazioni di probabilità $(P(F = 3900) + P(F = 3901) + ... + P(F = 4100))$.

- Proviamo comunque a ricorrere alla sola distribuzione binomiale e usiamo un calcolatore[31] per portare a termine il calcolo impostato già nel paragrafo 1.1. Risulta

$$P(3900 \leq F \leq 4100) \simeq 0,95877$$

I due numeri (quello trovato con la normale e quello trovato con la binomiale) sono "vicini" ma comunque diversi. Qual è quello "giusto"?
In realtà ... nessuno dei due. Infatti i due valori di probabilità sono stati ottenuti all'interno di due modelli diversi, binomiale e normale, che sono due *schematizzazioni* della situazione reale, tra le varie *possibili*. E tali modelli sono stati costruiti sulla base di nostre precise *scelte* sulla situazione reale: non sono la situazione reale. Concludiamo così che nessuno dei due valori è "più giusto" dell'altro e che entrambi rappresentano due possibili soluzioni del problema in esame.

[31] Con Excel si può usare la funzione $DISTRIB.BINOM.N(num\_successi; prove; probabilitá\_s; cumulativo)$ con cumulativo=VERO.

7.5 Applicazioni: dall'esame di Stato

Il TLC permette di risolvere in modo più agevole anche il quesito n.3 proposto nella simulazione della seconda prova[32] il 29 aprile 2016.

> Durante il picco massimo di un'epidemia di influenza, il 15% della popolazione è a casa ammalato:
>
> **a)** qual è la probabilità che in una classe di 20 alunni ce ne siano più di due assenti per influenza?
>
> **b)** descrivi le operazioni da compiere per verificare che, se l'intera scuola ha 500 alunni, la probabilità che ce ne siano più di 50 influenzati è maggiore del 99%.

a) *Decidiamo*[33] di modellizzare la situazione mediante la distribuzione binomiale. Consideriamo dunque la variabile aleatoria binomiale

$$S = \textit{numero di assenti per malattia}$$

I valori dei parametri della distribuzione sono:

$$\textit{numero di prove } \mathbf{n = 20}$$
$$\textit{probabilità che l'alunno si ammali } \mathbf{p = 0,15}.$$

Secondo il modello così costruito la probabilità dell'evento *"vi sono più di* 2 *assenti per malattia"* è[34]

$$\begin{aligned} P(S > 2) &= 1 - P(S \leq 2) \\ &= 1 - P(S = 2) - P(S = 1) - P(S = 0) \\ &= 1 - (190 \cdot 0,15^2 \cdot 0,85^{18} + 20 \cdot 0,15 \cdot 0,85^{19} + 0,85^{20}) \\ &\simeq \boxed{0,5951} \end{aligned}$$

b) Proviamo prima a schematizzare la situazione mediante un modello binomiale analogo al precedente. In questo caso i valori dei parametri della distribuzione sono:

$$\textit{numero di prove } \mathbf{n = 500}$$
$$\textit{probabilità che l'alunno si ammali } \mathbf{p = 0,15}.$$

Secondo tale modello la probabilità dell'evento *"vi sono più di* 50 *ammalati"* è:

$$\begin{aligned} P(S > 50) &= 1 - P(S \leq 50) \\ &= 1 - P(S = 50) - P(S = 49) - ... - P(S = 0) \end{aligned}$$

Il calcolo è dunque articolato.

[32] Tale quesito era già stato assegnato nella sessione straordinaria italiana di settembre 2015 e nella sessione ordinaria 2015, scuole italiane all'estero, Americhe.

[33] In effetti di decisione si tratta: sulla base delle informazioni fornite nel testo del quesito si possono costruire diversi **modelli**; la distribuzione binomiale è solo **uno** dei modelli possibili.

[34] Per la prima uguaglianza siamo passati all'evento complementare.

Per semplificarlo ricorriamo al TLC[35] e approssimiamo la distribuzione binomiale con la distribuzione normale che ha la stessa media μ e la stessa varianza σ^2 della binomiale considerata. Ossia

$$\mu = np = 75 \quad \text{e} \quad \sigma^2 = np(1-p) = 63,75$$

Detta X la variabile aleatoria normale *"numero di ammalati"*, vale

$$\mathbf{P(S > 50)} \underset{\substack{\uparrow \\ \textit{per il TLC}}}{\simeq} P(X > 50) \underset{\substack{\uparrow \\ \textit{standardizzando } X}}{=} P\left(Z > \tfrac{50-75}{\sqrt{63,75}}\right) \simeq P(Z > -3,13)$$

$$\underset{\substack{\uparrow \\ \textit{per simmetria}}}{=} P(Z < 3,13) \underset{\substack{\uparrow \\ \textit{ricorrendo alle} \\ \textit{tavole}}}{\simeq} \boxed{0,9991}$$

Abbiamo così verificato che la probabilità che ci siano più di 50 influenzati su una scuola di 500 alunni è maggiore del 99%.

[35] In realtà si poteva applicare il TLC anche nella situazione descritta nel punto a) del quesito. In tal caso, però, i vantaggi computazionali che comporta tale approccio non sono così evidenti come nella situazione prospettata nel punto b).