Azzolini Riccardo 2020-04-28

Applicazioni della distribuzione normale

1 Situazioni modellate da una distribuzione normale

Alcune situazioni tipicamente modellate da una distribuzione normale sono:

- Gli errori casuali nella misurazione di una grandezza fisica.

  La misura di una qualunque grandezza fisica può essere vista come la somma del valore esatto della grandezza (che è una costante) e dell'errore di misurazione, che è una variabile aleatoria X. La densità di X è solitamente una curva a campana, perché:

  - l'errore può essere per eccesso o per difetto, quindi X può assumere, in modi simmetrico, valori positivi o negativi;
  - l'errore è tende a essere abbastanza piccolo, o, meglio, errori più grandi sono meno probabili, quindi la curva decresce (rapidamente) man mano che ci si allontana dallo 0 (errore nullo), in entrambe le direzioni.

  Se l'errore presenta una componente sistematica, che si ripete a ogni misurazione, questa corrisponderà alla media μ della distribuzione. Altrimenti, la media sarà 0.

- La distribuzione di una caratteristica quantitativa di una popolazione, che presenta oscillazioni casuali attorno a una media.

  Molte grandezze (ad esempio statura, peso, ecc.) relative a una popolazione omogenea di persone possono essere rappresentate da una distribuzione gaussiana, avente μ uguale al valore medio della grandezza nella popolazione.

  Questo caso è simile al precedente, con la differenza che ciascuna misurazione è riferita a un individuo diverso, mentre prima si consideravano misurazioni ripetute di una stessa grandezza.

- La dimensione effettiva di oggetti prodotti in serie, che si cerca di produrre in modo identico.

  Qui il valore medio μ sarà la dimensione voluta, e la varianza σ^2 sarà la più piccola possibile.

  Ad esempio, se una ditta produce confezioni di biscotti da 250 g, il peso effettivo delle confezioni può essere rappresentato da una variabile aleatoria normale di valore medio $\mu = 250\,\text{g}$.

Rispetto al caso precedente, che analizzava una caratteristica "naturale" di una popolazione, la particolarità di questa situazione è che il processo produttivo ha proprio l'obiettivo di minimizzare la varianza risultante.

2 Campionamento

Le situazioni appena presentate sono problemi di **campionamento**. Gli elementi principali di questo tipo di problemi sono i seguenti:

- *Definizione*: Una **popolazione** è un insieme o collezione di oggetti, numeri, misure o osservazioni che sono oggetto di studio.
- *Definizione*: Un **campione** è una parte della popolazione, che viene selezionata per l'analisi.

In genere, è necessario trattare solo un campione della popolazione perché questa può essere molto numerosa, o addirittura infinita, e allora non è pratico / possibile considerarla tutta. Si impiegano allora tecniche di **statistica inferenziale** per trarre delle conclusioni sui parametri di una popolazione dai dati campionari (cioè relativi a un campione).

3 Problema: percentili

Problema: La variabile aleatoria X ha la distribuzione normale con valore medio $\mu = 19$ e varianza $\sigma^2 = 49$ (quindi $\sigma = \sqrt{49} = 7$); determinare i valori x_α tali che:

1. $P\{X > x_\alpha\} = 0.2 = 20\ \%$
2. $P\{X < x_\alpha\} = 0.9 = 90\ \%$

Soluzioni:

Come al solito, data l'impossibilità di effettuare calcoli analitici con la funzione di ripartizione della normale, bisogna consultare le tavole, dopo aver standardizzato la variabile e applicato eventuali simmetrie.

1. Per prima cosa si standardizza:

$$P\{X > x_\alpha\} = P\left\{Z > z_\alpha = \frac{x_\alpha - \mu}{\sigma}\right\} = P\left\{Z > z_\alpha = \frac{x_\alpha - 19}{7}\right\}$$

Poi, si passa all'evento complementare, che corrisponde alla funzione di ripartizione:

$$\begin{aligned} P\{Z > z_\alpha\} &= 1 - P\{Z < z_\alpha\} \\ 0.2 &= 1 - P\{Z < z_\alpha\} \\ P\{Z < z_\alpha\} &= 1 - 0.2 = 0.8 \end{aligned}$$

Le tavole riportano i valori a partire da $P\{Z < 0\} = 0.5$. Siccome qui si considera una probabilità uguale a $0.8 > 0.5$, essa sarà direttamente presente sulle tavole, e non è allora necessario applicare simmetrie.

Per determinare z_α, si individua nelle tavole il valore $F_Z(z)$ più vicino a 0.8, ad esempio 0.79955 (potrebbe variare a seconda delle specifiche tavole consultate), e si legge il valore di z corrispondente: $0.79955 = F_Z(0.84)$, ovvero

$$P\{Z < z_\alpha\} = 0.79955 \approx 0.8 \quad \text{per} \quad z_\alpha = 0.84$$

Infine, si calcola il valore di x_α corrispondente al valore standardizzato z_α:

$$\begin{aligned} z_\alpha &= \frac{x_\alpha - 19}{7} \\ x_\alpha &= 7z_\alpha + 19 = 7 \cdot 0.84 + 19 = 24.88 \end{aligned}$$

Quindi, ricapitolando, la soluzione è:

$$P\{X > x_\alpha\} = 0.2 \quad \text{per} \quad x_\alpha \approx 0.84$$

2. Come prima, si inizia standardizzando la variabile:

$$P\{X < x_\alpha\} = P\left\{Z < z_\alpha = \frac{x_\alpha - \mu}{\sigma}\right\} = P\left\{Z < z_\alpha = \frac{x_\alpha - 19}{7}\right\}$$

Siccome l'evento è già del tipo corrispondente alla funzione di ripartizione, e la probabilità richiesta è $0.9 > 0.5$, non è necessario passare all'evento complementare o applicare altre simmetrie, e si può invece passare direttamente alla lettura delle tavole, dalle quali si ricava che:

$$P\{Z < z_\alpha\} = 0.89973 \approx 0.9 \quad \text{per} \quad z_\alpha = 1.28$$

Ritornando alla variabile X non standardizzata:

$$\begin{aligned} x_\alpha &= 7z_\alpha + 19 = 7 \cdot 1.28 + 19 = 27.96 \\ P\{X < x_\alpha\} &= 0.9 \quad \text{per} \quad x_\alpha \approx 27.96 \end{aligned}$$

Nota: In generale, il valore x_α tale che $P\{X \leq x_\alpha\} = \alpha$ (come sempre, se X è continua, è equivalente considerare $X < x_\alpha$, senza l'uguale) si dice **quantile** di ordine α di X. In particolare, se α espresso come percentuale è un valore intero (ad esempio $\alpha = 0.03 = 3\ \%$, o $\alpha = 0.2 = 20\ \%$, ma non $\alpha = 0.125 = 12.5\ \%$), x_α è detto **percentile**.

4 Problema: parametri della distribuzione dai percentili

Nel problema precedente, si sono calcolati alcuni percentili di una variabile aleatoria normale a partire dai parametri della distribuzione. Viceversa, è anche possibile ricavare i parametri da alcuni percentili (quantili) noti: in particolare, se si conosce già la media o la deviazione standard / varianza, è sufficiente un percentile per determinare il parametro mancante, mentre se entrambi i parametri sono incogniti servono due percentili.

Problema: La variabile aleatoria X ha distribuzione normale con media μ e varianza σ^2 incognite. È noto che il 10 % dei valori di X è maggiore di 17.24, e che il 25 % dei valori è minore di 14.37. Trovare μ e σ^2 (o σ).

Soluzione:

Dal testo del problema, si sa che:

$$P\{X > 17.24\} = 0.1 \qquad P\{X < 14.37\} = 0.25$$

Standardizzando, ciò diventa:

$$P\left\{Z > \frac{17.24 - \mu}{\sigma}\right\} = 0.1 \qquad P\left\{Z < \frac{14.37 - \mu}{\sigma}\right\} = 0.25$$

- Considerando la prima di queste due probabilità note, e passando all'evento complementare per riportare il calcolo sulla funzione di ripartizione,

$$\begin{aligned} P\left\{Z > \frac{17.24 - \mu}{\sigma}\right\} &= 1 - P\left\{Z < \frac{17.24 - \mu}{\sigma}\right\} \\ 0.1 &= 1 - P\left\{Z < \frac{17.24 - \mu}{\sigma}\right\} \\ P\left\{Z < \frac{17.24 - \mu}{\sigma}\right\} &= 1 - 0.1 = 0.9 \end{aligned}$$

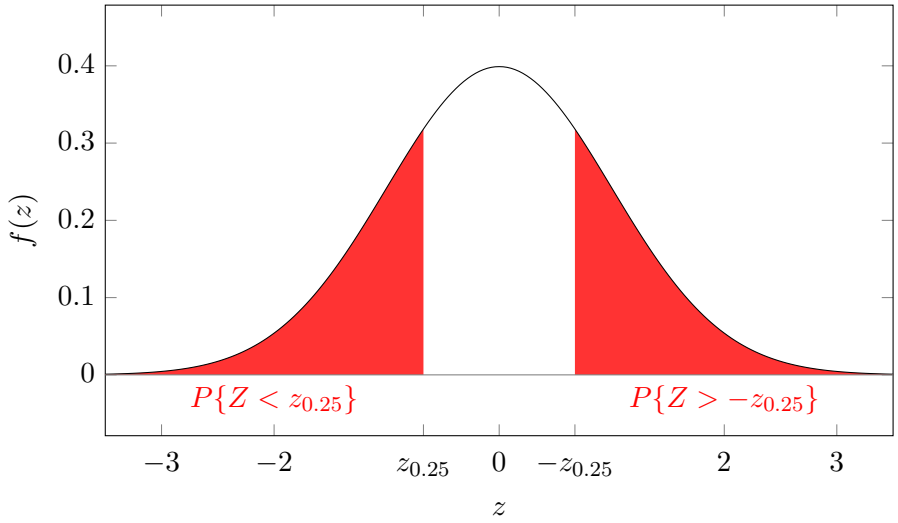
  si ricava dalle tavole che:

$$P\left\{Z < \frac{17.24 - \mu}{\sigma}\right\} = 0.89973 \approx 0.9 \quad \text{per} \quad \frac{17.24 - \mu}{\sigma} = 1.28$$

- La seconda probabilità nota,

$$P\left\{Z < z_{0.25} = \frac{14.37 - \mu}{\sigma}\right\} = 0.25$$

  è già espressa nella forma della funzione di ripartizione, ma il valore 0.25 non è in genere presente nelle tavole, che iniziano da $F_Z(0) = P\{Z < 0\} = 0.5$. Siccome la funzione di ripartizione è crescente, $F_Z(z_{0.25}) = 0.25 < 0.5$ implica che $z_{0.25} < 0$. Allora, $-z_{0.25}$ è il valore positivo simmetrico a $z_{0.25}$, e l'area sotto il grafico della densità alla destra di $-z_{0.25}$, ovvero $P\{Z > -z_{0.25}\}$, è uguale all'area a sinistra di $z_{0.25}$, cioè a $P\{Z < z_{0.25}\}$:



Adesso, passando all'evento complementare,

$$\begin{aligned} P\{Z > -z_{0.25}\} &= 1 - P\{Z < -z_{0.25}\} \\ 0.25 &= 1 - P\{Z < -z_{0.25}\} \\ P\{Z < -z_{0.25}\} &= 1 - 0.25 = 0.75 \end{aligned}$$

si arriva a un valore presente nelle tavole,

$$P\{Z < -z_{0.25}\} = 0.74857 \approx 0.75 \quad \text{per} \quad -z_{0.25} = 0.67$$

ricavando così l'equazione:

$$\begin{aligned} -z_{0.25} &= 0.67 \\ -\frac{14.37 - \mu}{\sigma} &= 0.67 \\ \frac{14.37 - \mu}{\sigma} &= -0.67 \end{aligned}$$

Infine, per trovare i valori di μ e σ, si mettono a sistema le equazioni ricavate da ciascuno dei due percentili noti:

$$\begin{cases} \dfrac{17.24 - \mu}{\sigma} = 1.28 \\ \dfrac{14.37 - \mu}{\sigma} = -0.67 \end{cases}$$

$$\begin{cases} 17.24 - \mu = 1.28\sigma \\ 14.37 - \mu = -0.67\sigma \end{cases}$$

$$\begin{cases} -\mu = 1.28\sigma - 17.24 \\ 14.37 = \mu - 0.67\sigma \end{cases}$$

$$
\begin{cases} \mu = 17.24 - 1.28\sigma \\ 14.37 = 17.24 - 1.28\sigma - 0.67\sigma \end{cases}
$$

$$
\begin{cases} \mu = 17.24 - 1.28\sigma \\ 14.37 - 17.24 = -1.95\sigma \end{cases}
$$

$$
\begin{cases} \mu = 17.24 - 1.28\sigma \\ 2.87 = 1.95\sigma \end{cases}
$$

$$
\begin{cases} \mu \approx 17.24 - 1.28 \cdot 1.47 \approx 17.24 - 1.88 = 15.36 \\ \sigma = \dfrac{2.87}{1.95} \approx 1.47 \end{cases}
$$

Allora, la variabile aleatoria X ha media $\mu \approx 15.36$ e deviazione standard $\sigma \approx 1.47$ (varianza $\sigma^2 \approx 1.47^2 \approx 2.16$).

5 Relazione tra la distribuzione binomiale e la normale

Sia X una variabile aleatoria binomiale, $X \sim B(n, p)$. Si ricorda che media e varianza della binomiale sono date dalle seguenti formule:

$$\mu = np \qquad \sigma^2 = np(1-p)$$

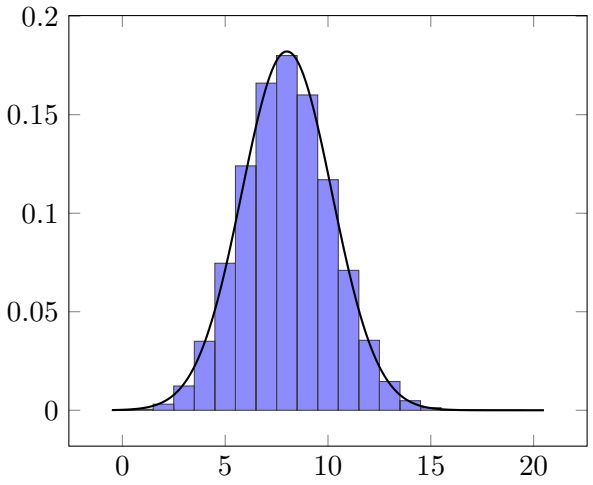
Quando n è grande e p è vicino a 0.5, la distribuzione di X può essere approssimata da una distribuzione normale avente la stessa media e la stessa varianza, cioè, standardizzando, da:

$$Z \approx \frac{X - np}{\sqrt{np(1-p)}}$$

L'approssimazione migliora al crescere di n (per $n \to +\infty$, si può dimostrare che le due distribuzioni coincidono); se p è più lontano da 0.5, si ritrova comunque una buona approssimazione all'aumentare di n. Come regola pratica, l'approssimazione è buona quando si verificano entrambe le condizioni $np \geq 5$ e $n(1-p) \geq 5$.[1] Ad esempio, per $n = 20$ e $p = 0.4$ (quindi $np = 8 \geq 5$ e $n(1-p) = 12 \geq 5$):

[1] Una sola di queste condizioni non basterebbe, perché, ad esempio, se si considerasse solo $np \geq 5$, anche valori di p molto vicini a 1 (lontani da 0.5) sembrerebbero dare una buona approssimazione già per n piccoli.

Densità binomiale $B(20, 0.4)$ e normale corrispondente



Quest'approssimazione è utile in quanto la densità binomiale per n grandi coinvolge grossi numeri fattoriali, e i calcoli sulla funzione di ripartizione possono comportare la somma di numerosi valori della densità, mentre con la distribuzione normale è sufficiente consultare le tavole.

5.1 Correzione di continuità

Per poter usare correttamente la distribuzione normale, che è continua, come approssimazione della distribuzione di una variabile aleatoria discreta (che, per semplicità, si suppone assuma valori interi), è necessario applicare la **correzione di continuità**: ogni valore x della variabile discreta viene rappresentato con l'intervallo di valori continui $(x - \frac{1}{2}, x + \frac{1}{2})$.

In altre parole, se X è una variabile aleatoria discreta e Y è una variabile aleatoria continua che la approssima, passando al continuo il calcolo di $P\{X = x\}$ deve essere sostituito da quello di

$$P\left\{x - \frac{1}{2} < Y < x + \frac{1}{2}\right\}$$

Ciò va fatto anche quando si considera un intervallo di valori di X:

- $P\{X \leq x\}$ diventa $P\left\{Y < x + \frac{1}{2}\right\}$
- $P\{X \geq x\}$ diventa $P\left\{Y > x - \frac{1}{2}\right\}$
- $P\{a \leq X \leq b\}$ diventa $P\left\{a - \frac{1}{2} < Y < b + \frac{1}{2}\right\}$

Senza questa correzione, invece, si osserva che, nel caso di un singolo valore ($\{X = x\}$), il calcolo darebbe sempre $P\{Y = x\} = 0$, perché la probabilità che una variabile aleatoria continua assuma esattamente uno specifico valore è appunto 0.

A livello intuitivo, la correzione di continuità corrisponde all'idea di considerare la probabilità dei valori continui in tutta la larghezza della barra dell'istogramma della distribuzione discreta, invece che solo il valore centrale (o il primo, o l'ultimo, ecc.).

6 Problema: lanci di un dado

Problema: Un dado viene lanciato 120 volte. Calcolare la probabilità che il numero 3 si presenti al più 15 volte.

Soluzione:

Sia X la variabile aleatoria discreta che conta il numero di lanci in cui esce 3. Essa ha distribuzione binomiale, $X \sim B(120, \frac{1}{6})$. Allora, il calcolo della soluzione del problema sarebbe

$$P\{X \leq 15\} = \sum_{k=0}^{15} \binom{120}{k} \left(\frac{1}{6}\right)^k \left(\frac{5}{6}\right)^{120-k}$$

che è molto laborioso: la somma ha 16 addendi, e ciascuno di essi comprende fattoriali e potenze grandi.

Conviene invece usare l'approssimazione con la distribuzione normale: siccome

$$np = 120 \cdot \frac{1}{6} = 20 \geq 5 \qquad n(1-p) = 120 \cdot \frac{5}{6} = 100 \geq 5$$

tale approssimazione è buona. Sia allora Y la variabile aleatoria normale, di parametri

$$\mu = np = 20 \qquad \sigma = \sqrt{np(1-p)} = \sqrt{120 \cdot \frac{1}{6} \cdot \frac{5}{6}} \approx \sqrt{16.67} \approx 4.08$$

che approssima X. Per la correzione di continuità, la probabilità

$$P\{X \leq 15\} = P\{0 \leq X \leq 15\}$$

corrisponde a

$$P\left\{-\frac{1}{2} < Y < 15 + \frac{1}{2}\right\} = P\{-0.5 < Y < 15.5\}$$

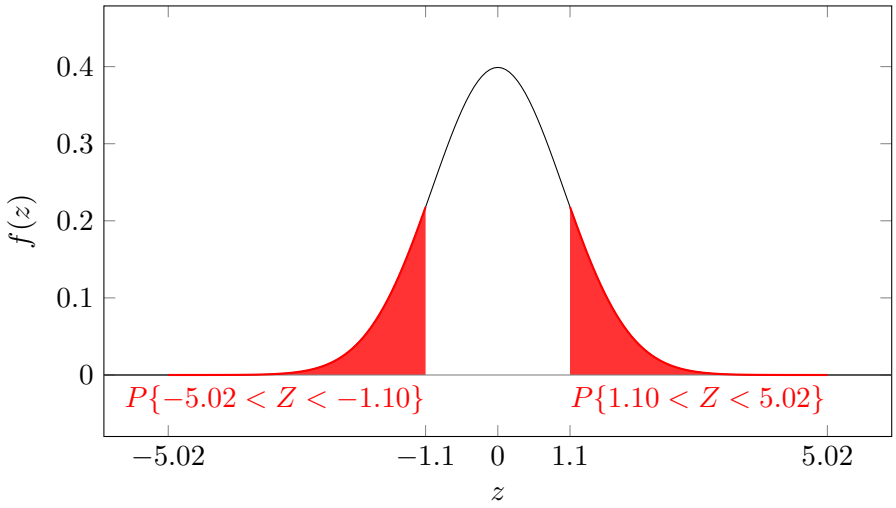
Standardizzando,

$$Z = \frac{Y - \mu}{\sigma} = \frac{Y - 20}{4.08}$$

la probabilità da calcolare diventa

$$P\{-0.5 < Y < 15.5\} = P\left\{\frac{-0.5-20}{4.08} < Z < \frac{15.5-20}{4.08}\right\} = P\{-5.02 < Z < -1.10\}$$

Gli estremi dell'intervallo sono negativi, quindi non presenti sulle tavole, ma, per la simmetria mostrata nel seguente grafico,



si ha che

$$\begin{aligned} P\{-5.02 < Z < -1.10\} &= P\{1.10 < Z < 5.02\} \\ &= P\{Z < 5.02\} - P\{Z < 1.10\} \\ &= F_Z(5.02) - F_Z(1.10) \\ &\approx 0.9999997 - 0.8643 \\ &\approx 0.1357 \end{aligned}$$

ovvero, ricapitolando, la probabilità che il numero 3 esca al più 15 volte è:

$$P\{X \leq 15\} \approx 0.1357 = 13.57\ \%$$

7 Problema: lanci di una moneta

Problema: Trovare la probabilità che, in 100 lanci di una moneta, testa si presenti 40 volte, usando la distribuzione normale per approssimare la distribuzione binomiale.

Soluzione:

Sia $X \sim B(100, \frac{1}{2})$ il numero di teste ottenute nei 100 lanci, e sia Y la variabile aleatoria normale che la approssima, avente parametri

$$\mu = np = 100 \cdot \frac{1}{2} = 50$$
$$\sigma = \sqrt{np(1-p)} = \sqrt{100 \cdot \frac{1}{2} \cdot \frac{1}{2}} = \sqrt{25} = 5$$

Siccome $n = 100$ è sufficientemente grande, e $p = 0.5$, qui l'approssimazione sarà molto buona.

La probabilità richiesta dal problema è $P\{X = 40\}$, che, con la correzione di continuità, diventa $P\{39.5 < Y < 40.5\}$. Quindi:

$$\begin{aligned}
P\{X = 40\} &\approx P\{39.5 < Y < 40.5\} \\
&= P\left\{\frac{39.5 - 50}{5} < Z < \frac{40.5 - 50}{5}\right\} && \text{(standardizzazione)} \\
&= P\{-2.1 < Z < -1.9\} \\
&= P\{1.9 < Z < 2.1\} && \text{(simmetria)} \\
&= F_Z(2.1) - F_Z(1.9) \\
&\approx 0.98214 - 0.97128 \\
&= 0.01086 = 1.086\ \%
\end{aligned}$$

8 Problema: chip di memoria

Problema: Il 20 % dei chip di memoria prodotti da un'azienda è difettoso; calcolare la probabilità che, in un campione di 100 chip scelto a caso per un controllo:

1. al più 15 siano difettosi;
2. esattamente 15 siano difettosi.

Soluzioni:

La variabile aleatoria X che conta il numero di chip difettosi nel campione segue la distribuzione binomiale $B(100, 0.2)$. Siccome

$$np = 100 \cdot 0.2 = 20 \geq 5 \qquad n(1-p) = 100 \cdot 0.8 = 80 \geq 5$$

si può approssimare X con una variabile Y normale di parametri

$$\mu = np = 20 \qquad \sigma = \sqrt{np(1-p)} = \sqrt{100 \cdot 0.2 \cdot 0.8} = \sqrt{16} = 4$$

I calcoli sono analoghi a quelli dei problemi precedenti, ricordando sempre di applicare la correzione di continuità:

$$\begin{aligned}
P\{X \leq 15\} &= P\{0 \leq X \leq 15\} \\
&\approx P\{-0.5 < Y < 15.5\} \\
&= P\left\{\frac{-0.5-20}{4} < Z < \frac{15.5-20}{4}\right\} \\
&= P\{-5.125 < Z < -1.125\} \\
&= P\{1.125 < Z < 5.125\} \\
&= F_Z(5.125) - F_Z(1.125) \\
&\approx 0.9999997 - 0.8708 \\
&\approx 0.1292 = 12.92\ \%
\end{aligned}$$

$$\begin{aligned}
P\{X = 15\} &\approx P\{14.5 < Y < 15.5\} \\
&= P\left\{\frac{14.5-20}{4} < Z < \frac{15.5-20}{4}\right\} \\
&= P\{-1.375 < Z < -1.125\} \\
&= P\{1.125 < Z < 1.375\} \\
&= F_Z(1.375) - F_Z(1.125) \\
&\approx 0.9162 - 0.8708 \\
&= 0.0454 = 4.54\ \%
\end{aligned}$$

9 Relazione tra la distribuzione di Poisson e la normale

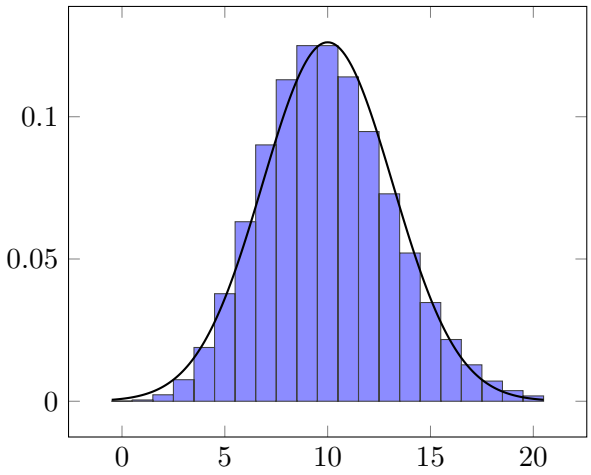
Il fatto che la distribuzione di Poisson sia il limite per $n \to +\infty$ di una distribuzione binomiale, e che la binomiale possa essere approssimata dalla normale, suggerisce che esista una relazione anche tra la Poisson e la normale.

Effettivamente, una variabile aleatoria X di Poisson di parametro λ può essere approssimata, per λ sufficientemente grandi, da una variabile aleatoria normale di parametri $\mu = \lambda$ e $\sigma^2 = \lambda$ (uguali alla media e alla varianza della Poisson), ovvero, standardizzando:

$$Z \approx \frac{X - \lambda}{\sqrt{\lambda}}$$

Come regola pratica, l'approssimazione è abbastanza buona per $\lambda \geq 10$.

Densità di Poisson e normale per $\lambda = 10$



Siccome anche la Poisson, come la binomiale, è una distribuzione discreta, per usare quest'approssimazione bisogna applicare la correzione di continuità.

10 Problema: incidenti

Problema: Il numero di incidenti d'auto che si verificano in un giorno a un incrocio è una variabile aleatoria con distribuzione di Poisson e media 1.4; calcolare la probabilità che accadano più di 50 incidenti in un periodo di 4 settimane.

Soluzione:

Il numero X di incidenti in 4 settimane, cioè 28 giorni, è dato dalla somma degli incidenti avvenuti in ciascuno dei 28 giorni:

$$X = X_1 + \cdots + X_{28} \quad \text{dove } X_i \sim \text{Poisson}(1.4)$$

Allora, per la regola della somma di Poisson (presentata in precedenza), si ha che

$$X \sim \text{Poisson}(28 \cdot 1.4) = \text{Poisson}(39.2)$$

Poiché $\lambda = 39.2 \geq 10$, si può approssimare X con una variabile aleatoria Y normale di parametri

$$\mu = \lambda = 39.2 \qquad \sigma = \sqrt{\lambda} = \sqrt{39.2}$$

Allora, la probabilità che accadano più di 50 incidenti si calcola come segue:

$$
\begin{aligned}
P\{X > 50\} &= P\{X \geq 51\} \\
&\approx P\{Y > 50.5\} && \text{(correzione di continuità)} \\
&= P\left\{ Z > \frac{50.5 - 39.2}{\sqrt{39.2}} \right\} && \text{(standardizzazione)} \\
&= P\{Z > 1.80\} \\
&= 1 - P\{Z < 1.80\} && \text{(evento complementare)} \\
&= 1 - F_Z(1.80) \\
&\approx 1 - 0.96407 \\
&= 0.03593 \approx 3.6\ \%
\end{aligned}
$$