Indici di posizione e dispersione per distribuzioni di variabili aleatorie

12 maggio 2017

Consideriamo i principali indici statistici che caratterizzano una distribuzione: indici di posizione, che forniscono informazioni del valore attorno al quale si posizionano i dati; b) indici di dispersione, che forniscono informazioni su quanto i dati si disperdano intorno ad un valore centrale

1 *Indici di posizione*

Aiutano a localizzare la distribuzione, ovvero individuare attorno a quale valore si concentra la distribuzione stessa. Sono espressi nella stessa unitá di misura della V.A.

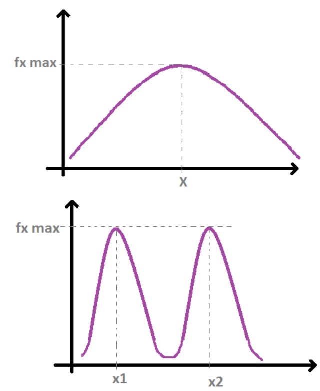

Figura 1: In alto, un esempio di distribuzione unimodale; in basso, un esempio di distribuzione multimodale

1.1 *Moda*

Data una distribuzione discreta $p_X(x)$ o continua $f_X(x)$, di una V.A. X la **moda** é il valore $X = x$ per cui la distribuzione é massima. Formalmente (caso continuo)

$$x_{moda} = \arg\max_x [f_X(x)].$$

Una distribuzione é detta *unimodale* se é presente un solo valore di massimo *multimodale* se sono presenti piú massimi. In quest'ultimo caso ciascun di essi si chiama valore modale.

1.2 *Mediana*

Data una distribuzione della V.A. X, la **mediana** é il valore $X = x_{mediana}$ che taglia in due parti equivalenti la distribuzione.

Formalmente:

$$P_X(\{X \leq x\}) = P_X(\{X > x\})$$
$$F_X(x) = 1 - F_X(x)$$
$$F_X(x) = \frac{1}{2}$$

Si noti che, in una distribuzione simmetrica, la moda e la mediana coincidono

1.3 *Media*

Data una distribuzione X, la **media** é visualizzabile come il *baricentro* della distribuzione. Formalmente, viene indicata come **valore atteso** oppure **speranza matematica** e si indica $< x >$, $E[X]$ (dove E sta per "expectation" ovvero speranza matematica), $\overline{X}$ oppure μ_X.

Nel caso discreto,

$$
\begin{aligned}
E[X] &= \sum x \cdot P_X(\{X = x\}) \\
&= \sum x \cdot p_X(x),
\end{aligned}
$$

dove la sommatoria é indicizzata da tutti i valori discreti che puó assumere la V.A. X.

Nel caso in cui X sia una V.A. continua, la sommatoria é sostituita da un integrale:

$$
E[X] = \int_{-\infty}^{+\infty} x \cdot f_X(x)dx
$$

La speranza matematica gode della proprietá di linearitá.

Prop. 1.1 *Se X é una variabile aleatoria degenere ovvero $X = \beta$ dove $\beta \in \mathbb{R}$ é una costante allora*

$$
E[\beta] = \beta
$$

Dimostrazione

$$
\begin{aligned}
E[\beta] &= \int_{-\infty}^{+\infty} \beta \cdot f_X(x)dx \\
&= \beta \cdot \int_{-\infty}^{+\infty} f_X(x)dx \\
&= \beta
\end{aligned}
$$

Prop. 1.2 *Dato $\alpha \in \mathbb{R}$*

$$
E[\alpha X] = \alpha \cdot E[X]
$$

Dimostrazione

$$
\begin{aligned}
E[\alpha X] &= \int_{-\infty}^{+\infty} \alpha \cdot x \cdot f_X(x)dx \\
&= \alpha \cdot \int_{-\infty}^{+\infty} x \cdot f_X(x)dx \\
&= \alpha \cdot E[X].
\end{aligned}
$$

Prop. 1.3 (Linearitá) *Dati $\alpha, \beta \in \mathbb{R}$*

$$
E[\alpha X + \beta] = \alpha \cdot E[X] + \beta
$$

Dimostrazione

$$
\begin{aligned}
E[\alpha X + \beta] &= \int_{-\infty}^{+\infty} (\alpha \cdot x + \beta) \cdot f_X(x) dx \\
&= \alpha \cdot \int_{-\infty}^{+\infty} x \cdot f_X(x) dx + \beta \cdot \int_{-\infty}^{+\infty} f_X(x) dx \\
&= \alpha \cdot E[X] + \beta.
\end{aligned}
$$

Si noti che la proprietá di linearitá (1.3) ci dice che se definiamo una nuova VA Y come funzione lineare di X

$$Y = \alpha X + \beta$$

allora il valore atteso di Y, cioé $E[Y]$, é funzione lineare di $E[X]$

$$E[Y] = \alpha E[X] + \beta$$

2 *Esempi con distribuzioni notevoli*

2.1 *Distribuzione Uniforme*

Consideriamo la distribuzione uniforme $Unif(a, b)$ la cui CDF é (cfr., Fig. 3):

$$F_X(x) = \frac{x-a}{b-a}.$$

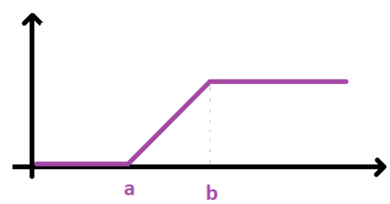

Figura 3: Funzione di ripartizione (CDF) della distribuzione uniforme $Unif(a, b)$

La mediana in si calcola come il valore $X = x$ tale che $F_X(x) = \frac{1}{2}$; pertanto:

$$
\begin{aligned}
\frac{x-a}{b-a} &= \frac{1}{2} \\
x - a &= \frac{b-a}{2} \\
x &= \frac{b-a}{2} + a \\
x_{mediana} &= \frac{b+a}{2}.
\end{aligned}
$$

Per quanto riguarda il valore atteso:

$$\begin{aligned}
E[X] &= \int_{-\infty}^{+\infty} x \cdot f_X(x)dx \\
&= \int_a^b x \cdot \frac{1}{b-a}dx \\
&= \frac{1}{b-a} \cdot [\frac{1}{2} \cdot x^2]_a^b \\
&= \frac{1}{2(b-a)} \cdot (b^2 - a^2) \\
&= \frac{1}{2} \cdot \frac{(b-a)(b+a)}{b-a} \\
&= \frac{b+a}{2}.
\end{aligned}$$

2.2 *Distribuzione Bernoulliana*

Nel caso della distribuzione Bernoulliana

$$Bern(p) = p^x \cdot q^{1-x},$$

il valore medio si calcola utilizzando la sommatoria con x che puó assumere solo due valori: 0 e 1.

$$\begin{aligned}
E[X] &= \sum_{x=0,1} x \cdot P_X(x) \\
&= \sum_{x=0,1} x \cdot p_X(x) \\
&= 0 \cdot p^0 \cdot q^1 + 1 \cdot p^1 \cdot q^0 \\
&= p
\end{aligned}$$

3 *Limiti degli indicatori di posizione*

In alcune distribuzioni particolari, alcuni indicatori di posizione non sono definiti ad eccezione della mediana: essa é sempre definita in qualunque distribuzione sia discreta che continua.

La moda non é sempre definita nelle distribuzioni continue, infatti nel caso in cui esista un intervallo di valori alla medesima frequenza massima, il valore modale non é definito. Un caso palese é rappresentato dalla distribuzione uniforme

Anche la media non é sempre definita: in alcuni casi l'integrale puó assumere valore divergente. Un esempio é fornito dalla densitá

$f_X(x) = \frac{1}{x^2}$:

$$\begin{aligned} E[X] &= \int_{-\infty}^{+\infty} x \cdot \frac{1}{x^2} dx \\ &= \int_{-\infty}^{+\infty} \frac{1}{x} dx \\ &= [\ln |x|]_{-\infty}^{+\infty} = \infty. \end{aligned}$$

4 *Indici di dispersione o ampiezza*

4.1 *Quartile*

Il quartile si puó considerare un'estensione della mediana: esso ripartisce la distribuzione in quattro parti equivalenti.

I quartili vengono indicati nel seguente modo: $q_{\frac{1}{4}}$(primo quartile), $q_{\frac{2}{4}}$ (secondo quartile, ovvero $q_{\frac{1}{2}}$ che corrisponde alla mediana), $q_{\frac{3}{4}}$(terzo quartile).

La quantitá

$$\Delta = q_{\frac{3}{4}} - q_{\frac{1}{4}}$$

é detta *scarto o intervallo interquartile*

Distribuzione uniforme Considerando ancora una VA distribuita uniformemente $X \sim Unif(a, b)$ con CDF

$$F_X(x) = \frac{x-a}{b-a},$$

i quartili si calcolano nel seguente modo.

Per il primo quartile:

$$x_{\frac{1}{4}} = a + \frac{b-a}{4}$$

Per il secondo quartile:

$$x_{\frac{1}{2}} = a + \frac{b-a}{2}$$

Per il terzo quartile:

$$x_{\frac{3}{4}} = a + \frac{3(b-a)}{4}$$

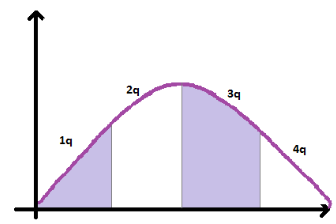

Figura 4: Rappresentazione grafica dei quartili

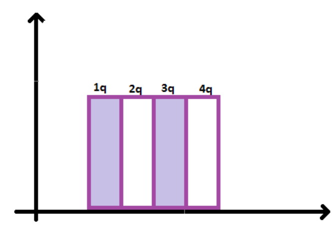

Figura 5: Quartili della distribuzione uniforme.

4.2 *Percentile*

I quartili si ottengono dividendo in quattro parti la distribuzione. Possiamo utilizzare una partizione piú fine, definendo i percentili. Il percentile ripartisce la distribuzione in cento parti equivalenti.

Definizione 4.1 *Sia $0 < p < 100$. Il **percentile di ordine** p di una distribuzione (o p-mo percentile, se p é intero) é il valore di x che delimita il primo p% dei dati dai rimanenti.*

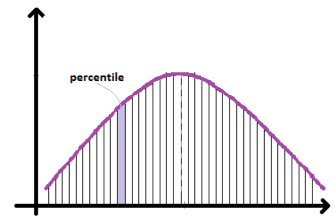

Figura 6: Rappresentazione grafica dei percentili.

Notiamo che, per esempio, $F_X(q_1) = \frac{1}{4} = 0.25$: dunque, il primo quartile corrisponde al 25-mo percentile. Analogamente, il secondo e il terzo corrispondono al 50-mo e al 75-mo

4.3 *Quantile*

Quartili e percentili sono casi particolari del quantile.

Definizione 4.2 *Data X V.A. e un valore α definito $0 < \alpha < 1$, il **quantile di ordine** α é il piú piccolo valore x_α tale che:*

$$P(X \leq x_\alpha) = F_X(x_\alpha) \geq \alpha$$

Nel caso continuo $F_X(x_\alpha) = \alpha$.

Si noti che la diseguaglianza $\geq$ é necessaria per contemplare anche il caso di VA discrete dove $F_X(x_\alpha)$ potrebbe non coincidere con α

Esempio 4.3 *Data la densitá*

$$f_X(x) = \frac{3x^2}{4^3} e^{-(\frac{x}{4})^3}$$

con $x \in (0, \infty)$ determinare l'intervallo I tale che la probabiliá di X sia compresa tra 0.3 e 0.9.

Soluzione.

Calcoliamo la CDF.

$$F_X(x) = \int_0^x \frac{3t^2}{4^3} e^{-(\frac{t}{4})^3} dt$$

Notiamo che:

$$\frac{3t^2}{4^3} e^{-(\frac{t}{4})^3} = -\frac{d}{dt}\left[e^{-(\frac{t}{4})^3}\right],$$

pertanto

$$\begin{aligned} F_X(x) &= \int_0^x \frac{3t^2}{4^3} e^{-(\frac{t}{4})^3} dt \\ &= -\int_0^x \frac{d}{dt}\left[e^{-(\frac{t}{4})^3}\right] dt \\ &= \left[-e^{-(\frac{t}{4})^3}\right]_0^x \\ &= 1 - e^{-(\frac{x}{4})^3} \end{aligned}$$

Troviamo l'estremo inferiore ponendo

$$F_X(x_{0.3}) = 1 - e^{-(\frac{x}{4})^3} = 0.3,$$

ovvero

$$e^{-(\frac{x}{4})^3 = 0.7},$$

da cui prendendo il logaritmo di entrambi i termini

$$-(\frac{x}{4})^3 = \ln 0.7 = -0.356$$

si ottiene $x_{0.3} \approx 2.84$.
Ripetendo gli stessi conti per $F_X(x_{0.9}) = 0.9$, *si ha che* $x_{0.9} = 5.28$.
In definitiva, l'intervallo

$$I = [2.84, 5.28]$$

é quello per cui

$$P(2.84 \leq X \leq 5.28) = 0.9 - 0.3 = 0.6$$

5 Indici di dispersione intorno alla media

Sia X una VA (discreta o continua), con valor medio

$$\langle X \rangle = E[X] = \begin{cases} \sum_x x \cdot p_X(x) & \text{se } X \text{ discreta,} \\ \int x \cdot f_X(x)\, dx & \text{se } X \text{ continua.} \end{cases} \tag{1}$$

Possiamo calcolare il valor medio di una funzione $g(X)$ definita sulla V.A. X:

$$\langle g(X) \rangle = E[g(X)] = \begin{cases} \sum_x g(x) \cdot p_X(x) & \text{se } X \text{ discreta,} \\ \int g(x) \cdot f_X(x)\, dx & \text{se } X \text{ continua.} \end{cases} \tag{2}$$

Definiamo una $g(X)$ come la funzione potenza r-ma di X, ovvero $g(X) = X^r$, che prende il nome di **momento di ordine** $\boldsymbol{r}$, dunque:

$$\langle X^r \rangle = E[X^r] = \begin{cases} \sum_x x^r \cdot p_X(x) & \text{se } X \text{ discreta,} \\ \int x^r \cdot f_X(x)\, dx & \text{se } X \text{ continua.} \end{cases} \tag{3}$$

Di particolare interesse é il **momento di ordine 2**, del quale vedremo il suo utilizzo piú avanti.

$$\left\langle X^2 \right\rangle = E[X^2] = \begin{cases} \sum_x x^2 \cdot p_X(x) & \text{se } X \text{ discreta,} \\ \int x^2 \cdot f_X(x)\, dx & \text{se } X \text{ continua.} \end{cases} \tag{4}$$

La **varianza** é un indice di **ampiezza** che identifica la **dispersione** di una VA rispetto al Valor Medio.

$$\sigma_X^2 = var(X) = \left\langle (X - E[X])^2 \right\rangle = E[(X - E[X])^2] = \begin{cases} \sum_x (x - \langle X \rangle)^2 \cdot p_X(x) & \text{se } X \text{ discreta,} \\ \int (x - \langle X \rangle)^2 \cdot f_X(x)\, dx & \text{se } X \text{ continua.} \end{cases} \tag{5}$$

La **deviazione standard** é definita come la radice quadrata della varianza, ed é anche nota come **scarto quadratico medio**.

$$\sigma_X = \sqrt{\sigma_X^2} = \sqrt{var(X)} \tag{6}$$

Nel misure sperimentali é utilizzata per identificare la precisione di una misura a fronte di numerose misurazioni:

$$\langle X \rangle \pm \sigma_X$$

Il raggio di un cilindro meccanico é di $172.0mm \pm 0.1mm$

Un modo alternativo per calcolare la varianza é fornito dal seguente.

Lemma 5.1 *La varianza di una V.A. equivale alla differenza fra il suo momento di ordine 2 e il quadrato della suo valore atteso:*

$$\sigma_X^2 = E[(X - E[X])^2] = E[X^2] - E[X]^2 \tag{7}$$

Dimostrazione

$$\begin{aligned}
E[(X - E[X])^2] &= \int (x - E[X])^2 \cdot f_X(x)\,dx = \\
&= \int (x^2 - 2x \cdot E[X] + E[X]^2) \cdot f_X(x)\,dx = \\
&= \underbrace{\int x^2 \cdot f_X(x)\,dx}_{E[X^2]} - 2E[X] \underbrace{\int x \cdot f_X(x)\,dx}_{E[X]} + E[X]^2 \underbrace{\int f_X(x)\,dx}_{1} = \\
&= E[X^2] - 2E[X]E[X] + E[X]^2 \cdot 1 = \\
&= E[X^2] - E[X]^2.
\end{aligned}$$

6 *Varianza e Deviazione Standard delle Variabili Aleatorie notevoli*

Vedremo ora il calcolo di questi indici di ampiezza per due distribuzioni notevoli: la Bernoulliana e l'Uniforme.

6.1 Varianza e deviazione standard di una VA Bernoulliana

$$\begin{aligned}
\sigma^2_{X\sim Bern(p)} &= E[X^2] - E[X]^2 = \\
&= \sum_{x=0,1} x^2 \cdot p_{X\sim Bern(p)}(x) - p^2 = \\
&= \sum_{x=0,1} x^2 \cdot p^x \cdot q^{1-x} - p^2 = \\
&= p - p^2 = p \cdot (1-p) = p \cdot q
\end{aligned} \tag{8}$$

$$\sigma_{X\sim Bern(p)} = \sqrt{p \cdot q}$$

6.2 Varianza e deviazione standard di una VA Uniforme

$$\begin{aligned}
\sigma^2_{X\sim Unif(a,b)} &= E[X^2] - E[X]^2 = \\
&= \int_a^b x^2 \cdot f_{X\sim Unif(a,b)}(x)\,dx - \left(\frac{a+b}{2}\right)^2 = \\
&= \int_a^b \frac{x^2}{b-a}\,dx - \frac{(a+b)^2}{4} = \\
&= \frac{1}{b-a}\int_a^b x^2\,dx - \frac{(a+b)^2}{4} = \\
&= \frac{1}{b-a}\left[\frac{x^3}{3}\right]_a^b - \frac{(a+b)^2}{4} = \\
&= \frac{b^3-a^3}{3(b-a)} - \frac{(a+b)^2}{4} = \\
&= \frac{4(b^3-a^3) - 3(b-a)(a+b)^2}{12(b-a)} = \\
&= \frac{(b-a)[4(b^2+ab+a^2) - 3(a^2+2ab+b^2)}{12(b-a)} = \\
&= \frac{4b^2+4ab+4a^2-3a^2-6ab-3b^2}{12} = \\
&= \frac{b^2-2ab+a^2}{12} = \\
&= \frac{(b-a)^2}{12}
\end{aligned} \tag{9}$$

$$\sigma_{X\sim Unif(a,b)} = \sqrt{\frac{(b-a)^2}{12}} = \frac{\sqrt{3}}{6} \cdot (b-a)$$

7 Disuguaglianza di Chebychev

Possiamo fornire un legame preciso fra varianza e deviazioni di X dal valore atteso $E[X]$ mediante la diseguaglianza di Chebychev

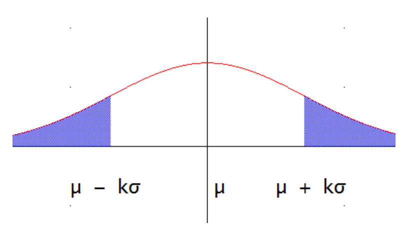

Figura 7: Rappresentazione della disuguaglianza di Chebychev. Si noti che, per costruzione, prendendo un valore x che cade nelle code $|x-\mu| \geq k\sigma$, ovvero la sua distanza da μ é $(x-\mu)^2 \geq k^2\sigma^2$

Teorema 7.1 *Sia X una VA, con $E[X] = \mu$ e $var(X) = \sigma^2$, allora $\forall k \in \mathbb{R} > 0$ vale la seguente:*

$$P_X(|X-\mu| \geq k\sigma) \leq \frac{1}{k^2} \tag{10}$$

o anche:

$$P_X(|X-\mu| < k\sigma) \geq 1 - \frac{1}{k^2}.$$

Dimostrazione

a) $0 < k < 1$.

Se $0 < k < 1$ *abbiamo*

$$P_X(|X-\mu| \geq k\sigma) \leq 1 < \tfrac{1}{k^2}$$

che é sempre valida per l'assioma di normalizzazione.

b) $k \geq 1$

Si consideri la Figura 7 e la diseguaglianza $(x-\mu)^2 \geq k^2\sigma^2$*. Per definizione di varianza e per l'ipotesi* $var(X) = \sigma^2$*:*

$$\begin{aligned}
\sigma^2 &= \int_{-\infty}^{+\infty} (x-\mu)^2 \cdot f(x)\,dx = \\
&= \int_{-\infty}^{\mu-k\sigma} (x-\mu)^2 \cdot f(x)\,dx + \int_{\mu-k\sigma}^{\mu+k\sigma} (x-\mu)^2 \cdot f(x)\,dx + \int_{\mu+k\sigma}^{+\infty} (x-\mu)^2 \cdot f(x)\,dx \geq \\
&\geq k^2\sigma^2 \int_{-\infty}^{\mu-k\sigma} f(x)\,dx + \int_{\mu-k\sigma}^{\mu+k\sigma} (x-\mu)^2 f(x)\,dx + k^2\sigma^2 \int_{\mu+k\sigma}^{+\infty} f(x)\,dx \geq \\
&\geq k^2\sigma^2 \left(\underbrace{\int_{-\infty}^{\mu-k\sigma} f(x)\,dx}_{P_X(X\leq\mu-k\sigma)} + \underbrace{\int_{\mu+k\sigma}^{+\infty} f(x)\,dx}_{P_X(X\geq\mu+k\sigma)} \right)
\end{aligned}$$

$$\Rightarrow \sigma^2 \geq k^2\sigma^2 \cdot P_X(|X-\mu| \geq k\sigma)$$

$$\Rightarrow \frac{1}{k^2} \geq P_X(|X-\mu| \geq k\sigma)$$

Questa é una relazione del tutto generale, dimostrabile anche nel caso discreto: l'unica condizione é l'esistenza di media e varianza finite. La diseguaglianza ci dice che la probabilitá che X ha di deviare di $k\sigma$ dal suo valore atteso tende rapidamente a 0 per $k \to \infty$.

In pratica ci dice che negli intervalli centrati sulla media e ampi 2σ e 3σ sono compresi rispettivamente *almeno* il 75% e il 90% della probabilitá totale. Una Tabella precisa che indica, al variare di k, la percentuale di probabilitá totale intorno alla media e nelle code destra e sinistra é riportata in Figura 8.

La diseguaglianza dá luogo alla **legge 3σ generalizzata**, che con-

Legge 3σ generalizzata

siste nel ritenere trascurabili *per qualunque distribuzione statistica*, le probabilitá di avere valori fuori dall'intervallo

$$[\mu - 3\sigma, \mu + 3\sigma]$$

k	Min. % within *k* standard deviations of mean	Max. % beyond *k* standard deviations from mean
1	0%	100%
$\sqrt{2}$	50%	50%
1.5	55.56%	44.44%
2	75%	25%
3	88.8889%	11.1111%
4	93.75%	6.25%
5	96%	4%
6	97.2222%	2.7778%
7	97.9592%	2.0408%
8	98.4375%	1.5625%
9	98.7654%	1.2346%
10	99%	1%

Figura 8: Percentuale di probabilitá totale nell'intorno $k\sigma$ della media e nelle code destra e sinistra

Esempio 7.2 *Il numero di pezzi prodotti da una fabbrica durante una settimana é una VA di media $\mu = 50$ e di varianza pari a $\sigma^2 = 25$. Cosa si puó dire sulla probabilitá che la produzione sia compresa fra i 40 e 60 pezzi?*

Soluzione. *Possiamo ragionare in due modi. Poiché $\sigma^2 = 25$ allora $\sigma = 5$; ovvero, la produzione é caratterizzabile come di*

$$50 \pm 5$$

pezzi a settimana. Affinché sia compresa fra i 40 e 60 pezzi occorre che

$$50 \pm 2\sigma$$

da cui $k = 2$. Applicando Chebychev, nella forma $P_X(|X - \mu| \leq k\sigma) \geq 1 - \frac{1}{k^2}$ si ha che

$$P_X(40 \leq X \leq 60) = P_X(|X - 50| \leq 10) \geq 1 - \frac{1}{4} = \frac{3}{4}$$

Guardando la tabella di Figura 8, la probabilitá che la produzione sia compresa fra i 40 *e* 60 *pezzi é almeno del* 75%.

Oppure, ponendo $r = k\sigma$ *in Chebychev, la riscrivo come*

$$P_X(|X-\mu|\geq k\sigma) \leq \frac{1}{k^2} = P_X(|X-\mu|\geq r) \leq \frac{\sigma^2}{r^2}$$

da cui direttamente:

$$P_X(|X-\mu|\geq r) \leq \frac{\sigma^2}{r^2} = P_X(|X-50|\geq 10) \leq \frac{25}{10^2} = \frac{1}{4}$$

.

Dunque $P_X(40 \leq X \leq 60) = 1 - \frac{1}{4}$.

Si noti ancora una volta che il risultato é stato ottenuto senza avere identificato un modello probabilistico preciso per la distribuzione $P_X(X = x)$