Esercitazione 2 - Statistica (parte II)

Davide Passaretti

16/2/2017

Contents

1 Variabili casuali continue

Una variabile casuale continua può assumere un'infinità di valori. Per questo motivo, la probabilità che tale v.c. assuma un determinato valore è zero (evento *quasi impossibile*). Esempio: qual è la probabilità che un individuo sia alto esattamente 1.7097649 metri? Ovviamente zero.

Per le v.c. continue abbiamo una `pdf` o *funzione di densità*, che è l'equivalente della `pmf` che abbiamo già visto nelle v.c. discrete. Tuttavia, noi lavoreremo sulla `cdf` o *funzione di ripartizione*, che cumula la densità da $-\infty$ fino al punto d'interesse. Grazie alla funzione di ripartizione potremo ottenere la probabilità in un intervallo per le v.c. che considereremo.

2 La variabile casuale Uniforme Continua

La varaibile casuale Uniforme Continua attribuisce la stessa probabilità a tutti i punti dell'intervallo su cui è definita e zero altrove. Dato un intervallo chiuso $[a, b]$, la v.c. Uniforme:

$$\mathcal{U} \sim \mathbf{Unif}(a, b)$$

ha `pdf` (funzione di densità) uguale a:

$$f_{\mathcal{U}}(x; a, b) = \frac{1}{b-a} \text{ per } a \leq x \leq b$$

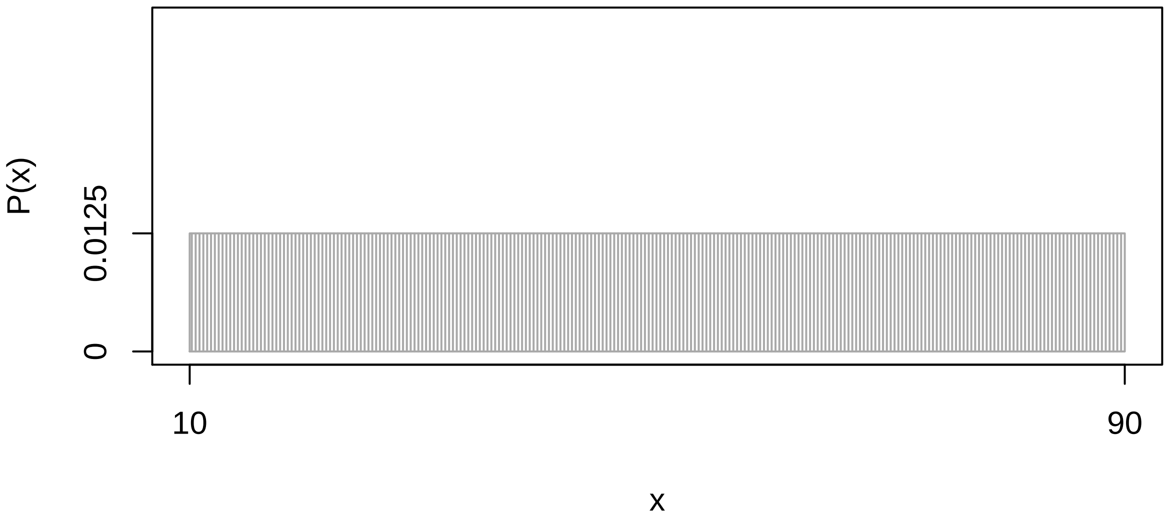
e `cdf` (funzione di ripartizione) uguale a:

$$F_{\mathcal{U}}(x; a, b) = \mathbf{P}(\mathcal{U} \leq x) = \int_0^x f_{\mathcal{U}}(t; a, b)\, dt = \frac{x-a}{b-a} \text{ per } a \leq x \leq b$$

L'area di probabilità in un intervallo (x_1, x_2), con $a \leq x_1 \leq x_2 \leq b$, è determinabile in modo veloce, in quanto *la probabilità nella v.c. Uniforme è proporzionale all'ampiezza dell'intervallo considerato*:

$$\mathbf{P}(x_1 \leq \mathcal{U} \leq x_2) = F_{\mathcal{U}}(x_2; a, b) - F_{\mathcal{U}}(x_1; a, b) = \frac{x_2 - a}{b - a} - \frac{x_1 - a}{b - a} = \frac{x_2 - x_1}{b - a}$$

Facciamo qualche esercizio. Supponiamo che Giovanni possa timbrare il cartellino a lavoro dalle 8.10 alle 9.30 e che il suo orario di entrata si distribuisca uniformemente nell'arco di questi 80 minuti. Quindi potremmo formare un intervallo in cui $a = 10$ (10 minuti dopo le 8.00) e $b = 90$ (90 minuti dopo le 8.00). Il valore della `pdf` per ciascun punto dell'intervallo è $\frac{1}{80} = 0.0125$, mentre al di fuori dell'intervallo la probabilità è zero.



a) Qual è la probabilità che Giovanni timbri il cartellino alle 8.20?

Trattandosi di v.c. continua, la sua probabilità in un punto è zero.

b) Qual è la probabilità che Giovanni timbri il cartellino entro le 8.20?

$$\mathbf{P}(\mathcal{U} \leq 20) = \frac{20 - 10}{80} = \frac{1}{8} = 0.125$$

c) Qual è la probabilità che Giovanni timbri il cartellino tra le 8.25 e le 8.35?

$$\mathbf{P}(25 \leq \mathcal{U} \leq 35) = \frac{35 - 25}{80} = \frac{1}{8} = 0.125$$

La probabilità è identica a quella del punto precedente in quanto il tempo che intercorre tra le 8.10 e le 8.20 è lo stesso che intercorre tra le 8.25 e le 8.35, cioè 10 minuti. La probabilità trovata non è altro che l'area associata ad un qualunque intervallo di 10 minuti compreso in $[a, b]$.

d) Poniamo il caso che alle 9.00 Giovanni non sia ancora arrivato. Qual è la probabilità che Giovanni timbri il cartellino tra le 9.10 e le 9.15?

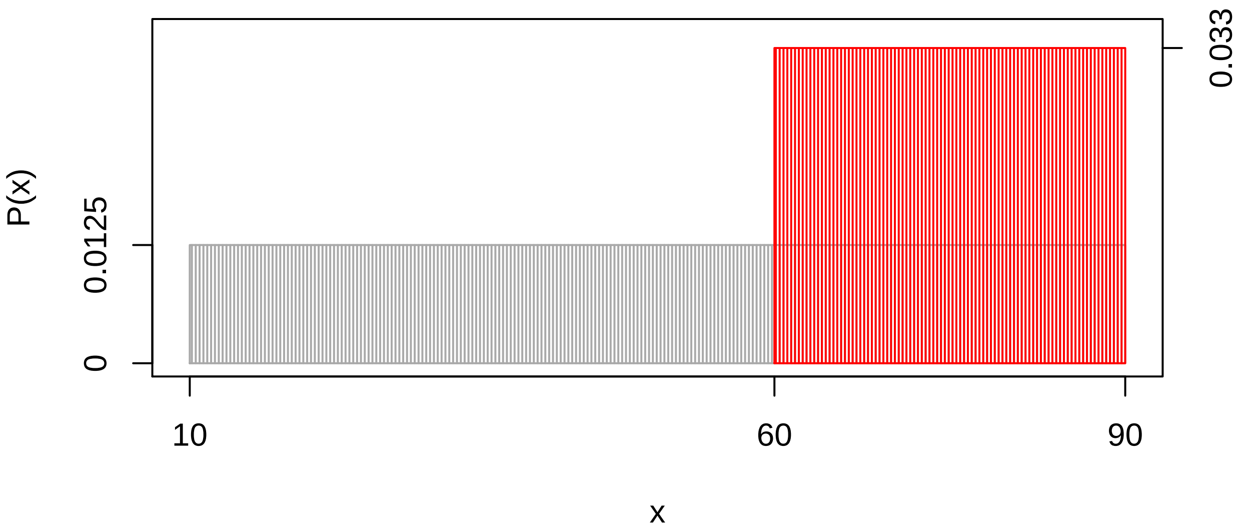
Il tempo rimasto fino alle 9.30 (cioè mezz'ora) è il nuovo intervallo di riferimento $[a', b]$, dove $a' = 60$:

$$\mathcal{U}' \sim \mathbf{Unif}(60, 90)$$

Qui siamo interessati alla probabilità tra le 9.10 e le 9.15, cioè in un intervallo di 5 minuti:

$$\mathbf{P}(70 \leq \mathcal{U}' \leq 75) = \frac{5}{30} = 0.16667$$

La figura seguente mostra la nuova area di probabilità (in rosso) su cui abbiamo ragionato.



3 La variabile casuale Esponenziale Negativa

La variabile casuale Esponenziale Negativa descrive la "durata di vita" di un fenomeno che non invecchia (quindi vi è *assenza di memoria*). Il suo unico parametro è λ, che è definito in R^+ e coincide con il reciproco sia del valore atteso e che dello scarto quadratico medio della variabile:

$$\mathcal{E} \sim \mathbf{Exp}(\lambda); \quad E(\mathcal{E}) = \frac{1}{\lambda}; \quad Var(\mathcal{E}) = \frac{1}{\lambda^2}$$

La `pdf` (funzione di densità) è:

$$f_{\mathcal{E}}(x; \lambda) = \lambda e^{-\lambda x}$$

La `cdf` (funzione di ripartizione) è:

$$F_{\mathcal{E}}(x; \lambda) = \mathbf{P}(\mathcal{E} \leq x) = \int_0^x f_{\mathcal{E}}(t; \lambda)\, dt = 1 - e^{-\lambda x}$$

Si può interpretare λ come un *rate parameter* (parametro di frequenza): se una batteria dura in media 2 anni e mezzo, allora $\lambda = \frac{1}{2.5}$ puà essere letta come "mediamente, bisogna cambiare batteria con una frequenza di una volta ogni 2 anni e mezzo".

In ambito scientifico, la distribuzione esponenziale si usa molto per descrivere fenomeni legati alla radioattività. Per esempio, il tempo di decadimento di un isotopo radioattivo viene solitamente modellato in funzione della sua vita media tramite la distribuzione esponenziale $\mathcal{E}(\lambda = \frac{1}{\tau})$, dove λ è detta costante di decadimento, mentre la speranza di vita è ovviamente il valore atteso, cioè τ.

a) Qual è la probabilità che l'isotopo duri meno della sua durata media (speranza di vita)?

$$\mathbf{P}(\mathcal{E} \leq \tau) = 1 - e^{-\frac{1}{\tau} \times \tau} = 1 - e^{-1} = 0.6321206$$

b) Qual è la probabilità che l'isotopo duri più del triplo della sua speranza di vita?

Ciò è determinabile calcolando il complemento a 1 della funzione di ripartizione:

$$\mathbf{P}(\mathcal{E} \geq 3\tau) = 1 - F_{\mathcal{E}}\left(3\tau; \frac{1}{\tau}\right) = 1 - \left(1 - e^{-\frac{1}{\tau}\times 3\tau}\right) = e^{-3} = 0.04978707$$

c) Supponiamo che dopo un tempo t, l'isotopo non sia ancora decaduto. Quanto tempo in media si dovrà ancora aspettare prima che l'isotopo decada?

Essendo la v.c. Esponenziale Negativa priva di memoria, non ha senso considerare il fatto che sia già trascorso del tempo: dal momento che l'isotopo non è ancora decaduto, ci troviamo di nuovo nelle condizioni iniziali (cioè l'isotopo *non è invecchiato*). Per questo motivo, dovremo aspettare in media ancora un tempo uguale a τ affinché l'isotopo decada.

d) Qual è la probabilità che si debba aspettare esattamente un tempo uguale a τ affinché l'isotopo decada?

Tale probabilità è uguale a 0, trattandosi di v.c. continua.

3.1 Legame tra la v.c. Esponenziale Negativa e la v.c. di Poisson

La v.c. di Poisson di parametro λ descrive il numero di eventi (rari e tra loro indipendenti) che si verificano in un intervallo temporale e spaziale. Il tempo che intercorre tra un evento indipendente e l'altro è descritto da una v.c. Esponenziale Negativa avente il medesimo λ.

Per esempio, il numero di auto che transitano in una zona rurale in un'ora è descritto da una legge di Poisson di parametro $\lambda = 2$. Ciò significa che il tempo che intercorre tra il passaggio di due auto è descritto da una v.c. Esponenziale Negativa che ha valore atteso uguale a $\frac{1}{2}$: bisogna aspettare in media mezz'ora prima di vedere un'auto.

a) Qual è la probabilità che transitino più di due auto in un'ora?

Qui va applicata la Poisson di parametro $\lambda = 2$, con $k > 2$. Quindi, basta calcolare il complemento ad 1 della probabilità di avere un numero di auto minore o uguale a 2.

$$\mathbf{P}(\mathcal{P} \geq 2) = 1 - [(\mathcal{P} = 0) + (\mathcal{P} = 1) + (\mathcal{P} = 2)] = 1 - \left[\frac{2^0\, e^{-2}}{0!} + \frac{2^1\, e^{-2}}{1!} + \frac{2^2\, e^{-2}}{2!}\right] = 1 - 5\, e^{-2} = 0.3233236$$

b) Qual è la probabilità di dover aspettare più di mezz'ora per veder transitare un'auto?

Qui va applicata la v.c. Esponenziale di parametro $\lambda = 2$, con $x \geq \frac{1}{2}$, cioè $x \geq E(\mathcal{E})$.

$$\mathbf{P}\left(\mathcal{E} \geq \frac{1}{2}\right) = e^{-2\times\frac{1}{2}} = e^{-1} = 0.3678794$$

c) Qual è la probabilità di dover aspettare dai 10 ai 20 minuti per veder transitare un'auto?

Dieci minuti e venti minuti sono rispettivamente $\frac{1}{6}$ e $\frac{1}{3}$ di un'ora. La probabiltà finale si ottiene come differenza:

$$\mathbf{P}\left(\frac{1}{6} \leq \mathcal{E} \leq \frac{1}{3}\right) = F_{\mathcal{E}}\left(\frac{1}{3}; 2\right) - F_{\mathcal{E}}\left(\frac{1}{6}; 2\right) = \left(1 - e^{-2\times\frac{1}{3}}\right) - \left(1 - e^{-2\times\frac{1}{6}}\right) = e^{-\frac{1}{3}} - e^{-\frac{2}{3}} = 0.2031142$$

d) Qual è la probabilità che transiti più di un'auto in mezz'ora?

Qui va applicata la Poisson di parametro $\lambda' = 1$. Tale parametro conta il numero medio di auto in mezz'ora ed è l'equivalente di $\lambda = 2$ (2 auto in un'ora). Ovviamente, in questo caso vogliamo $k > 1$.

$$\mathbf{P}(\mathcal{P}' \geq 1) = 1 - [(\mathcal{P}' = 0) + (\mathcal{P}' = 1)] = 1 - \left[\frac{1^0\, e^{-1}}{0!} + \frac{1^1\, e^{-1}}{1!}\right] = 1 - 2\, e^{-1} = 0.2642411$$

4 La variabile casuale Normale

La variabile casuale Normale (o Gaussiana) è la più comune v.c. in Statistica. È contraddistinta da una forma campanulare, dunque **simmetrica**. Essa descrive fenomeni i cui valori più probabili si ammassano attorno alla media, mentre man mano che ci si allontana dalla media – verso sinistra o verso destra – si trovano valori sempre meno probabili. La sua vasta diffusione si deve anche al **T**eorema del **L**imite **C**entrale, secondo il quale una combinazione lineare di variabili casuali *i.i.d.* si distribuisce asintoticamente come una Normale. Inoltre, c'è da sottolineare che una qualunque combinazione lineare di v.c. Normali è distribuita normalmente (*proprietà riproduttiva* della Normale).

La v.c. Normale ha due parametri: la media μ, che coincide con mediana e moda a causa della simmetria, e la varianza σ^2, che influisce sull'ampiezza della "campana":

$$X \sim \mathbf{N}(\mu, \sigma^2)$$

La `pdf` è la seguente:

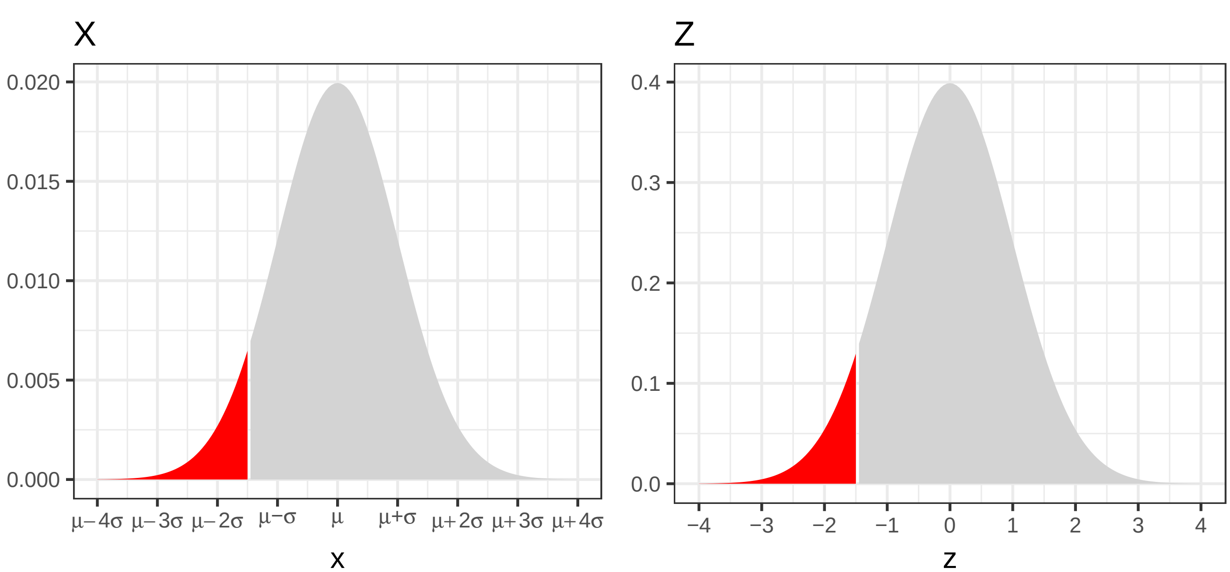
$$f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

La `cdf`, invece, non è determinabile analiticamente. Per questo, ci serviremo della tavola. Per avvalerci di tale tavola, dobbiamo prima standardizzare X, cioè centrarla e poi dividerla per il suo scarto quadratico medio:

$$\mathcal{Z} = \frac{X - \mu}{\sigma} \sim \mathbf{N}(0, 1)$$

La standardizzazione può essere sfruttata in quanto è possibile dimostrare che:

$$\underbrace{\mathbf{P}(X \leq x)}_{\texttt{cdf di } X \text{ in } x} = \underbrace{\mathbf{P}\left(\mathcal{Z} \leq \frac{x-\mu}{\sigma}\right)}_{\texttt{cdf di } \mathcal{Z} \text{ in } z = \frac{x-\mu}{\sigma}} = \Phi(z)$$



Supponiamo che le persone presenti quotidianamente in una scuola abbiano un peso corporeo che si distribuisce come una Normale con media 72 Kg e scarto quadratico medio di 20 Kg:

$$X \sim \mathbf{N}(\mu = 72, \sigma^2 = 400)$$

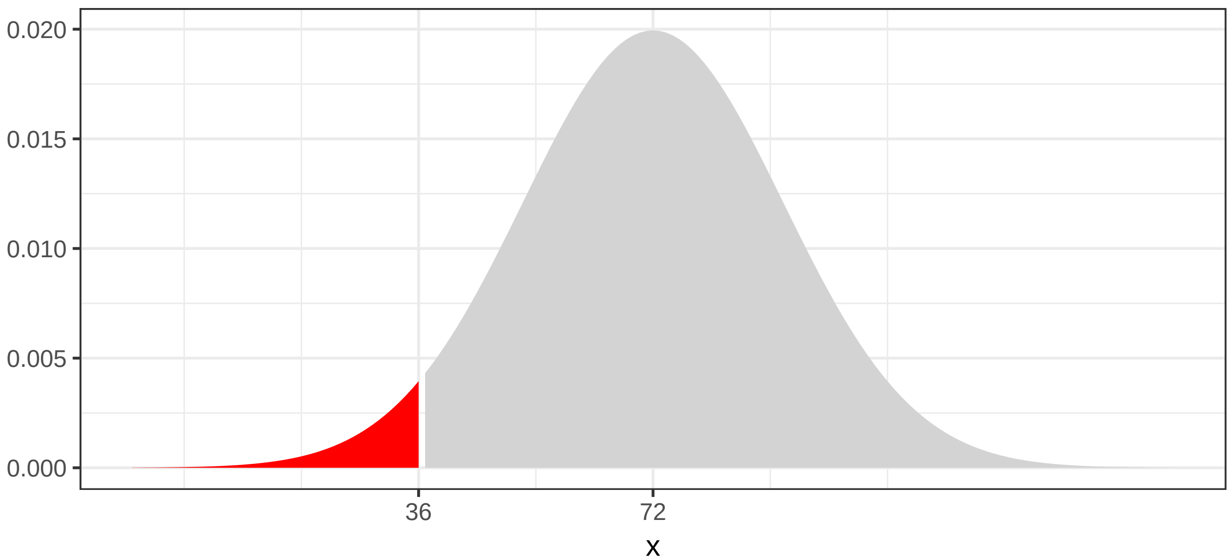
a) Qual è la probabilità di ottenere un peso pari al peso medio?

Trattandosi di v.c. continua, la probabilità in un punto è zero.

b) Qual è la probabilità di ottenere un peso inferiore al peso medio? E quella di ottenere un peso superiore al peso medio?

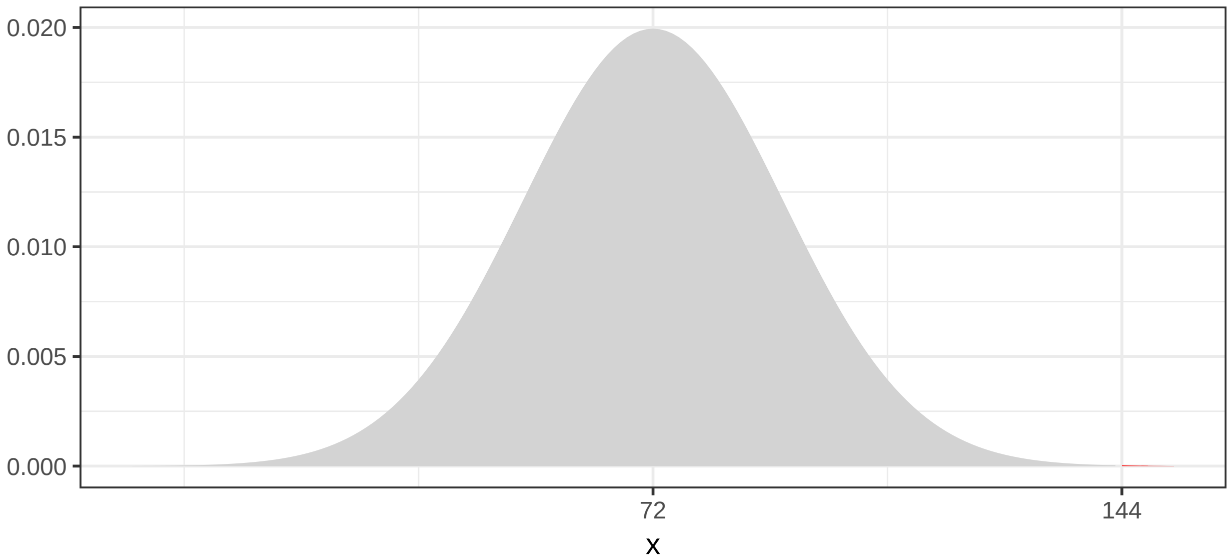
Essendo la media uguale alla mediana, essa divide esattamente in due la "campana": la probabilità di un peso inferiore alla media è uguale alla probabilità di un peso superiore alla media, cioè sono entrambe 0.5.

c) Qual è la probabilità di ottenere un peso minore della metà del peso medio?



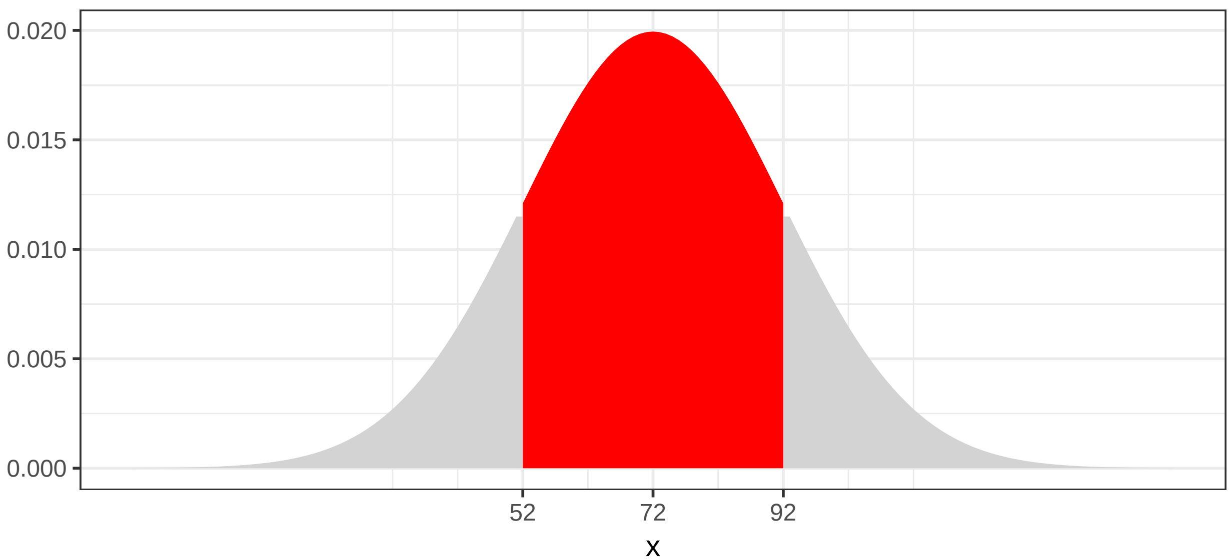
$$\mathbf{P}\left(X \leq \frac{72}{2}\right) = \mathbf{P}\left(\mathcal{Z} \leq \frac{36-72}{20}\right) = \mathbf{P}(\mathcal{Z} \leq -1.8) = \mathbf{P}(\mathcal{Z} \geq 1.8) = 1 - \underbrace{\mathbf{P}(\mathcal{Z} \leq 1.8)}_{\Phi(1.8)} = 1 - 0.9641 = 0.0359$$

d) Qual è la probabilità di ottenere un peso maggiore del doppio del peso medio?



$$\mathbf{P}(X \geq 72 \times 2) = 1 - \mathbf{P}\left(\mathcal{Z} \leq \frac{144-72}{20}\right) = 1 - \Phi(3.6) = 1 - 0.9998 = 0.0002$$

e) Qual è la probabilità di ottenere un peso tra 52 e 92 Kg?



$$\mathbf{P}(52 \leq X \leq 92) = \mathbf{P}\left(\mathcal{Z} \leq \frac{92-72}{20}\right) - \mathbf{P}\left(\mathcal{Z} \leq \frac{52-72}{20}\right) = \mathbf{P}(\mathcal{Z} \leq 1) - \mathbf{P}(\mathcal{Z} \leq -1) =$$

$$= \mathbf{P}(\mathcal{Z} \leq 1) - \mathbf{P}(\mathcal{Z} \geq 1) = \mathbf{P}(\mathcal{Z} \leq 1) - [1 - \mathbf{P}(\mathcal{Z} \leq 1)] = 2 \times \Phi(1) - 1 = 2 \times 0.8413 - 1 = 0.6826$$

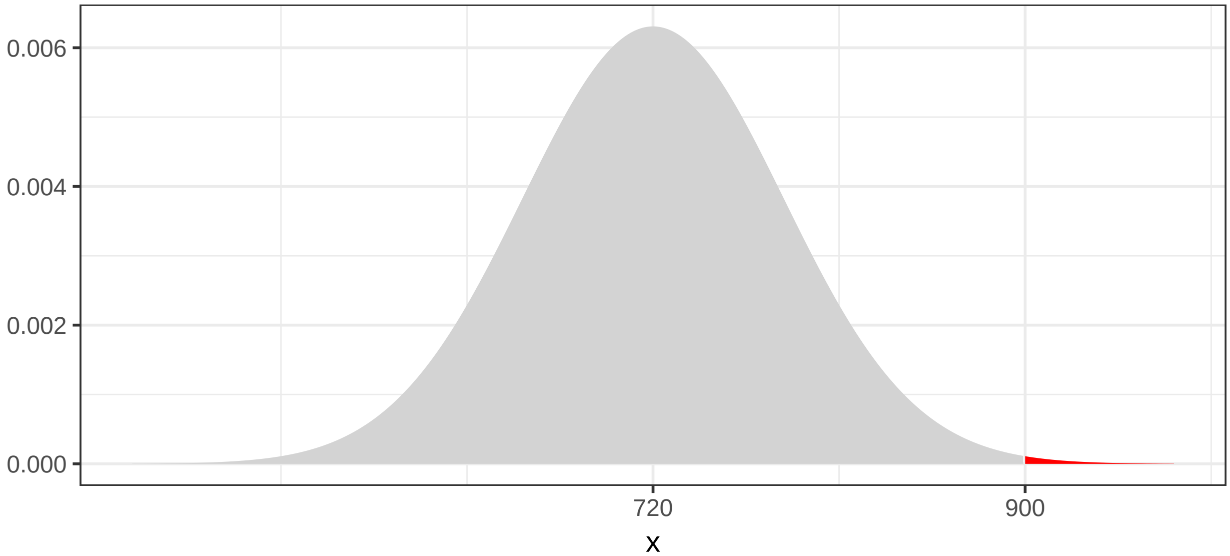
Supponiamo che questa scuola abbia un solo ascensore con capienza 10 persone e che tale ascensore si riempia sempre ad ogni salita o discesa. Il preside è preoccupato riguardo il peso che l'ascensore può sostenere. Il limite per essere in piena sicurezza è 900 Kg.

f) Come si distribuisce tale peso totale, assumendo che i pesi siano tra loro indipendenti?

Si tratta della v.c. somma di 10 v.c. Normali indipendenti con stessa media e varianza:

$$X_{tot} \sim \mathbf{N}(\mu_{tot} = n \times \mu,\, \sigma^2_{tot} = n \times \sigma^2) = \mathbf{N}(\mu_{tot} = 720,\, \sigma^2_{tot} = 4000)$$

g) Qual è la probabilità che la somma dei pesi delle 10 persone in ascensore superi il limite?



$$\mathbf{P}(X_{tot} \geq 900) = 1 - \mathbf{P}\left(\mathcal{Z}_{tot} \leq \frac{900 - 720}{\sqrt{4000}}\right) = 1 - \Phi_{tot}(2.85) = 1 - 0.9978 = 0.0022$$

h) Qual è la probabilità di ottenere almeno un peso totale superiore al limite in 500 viaggi indipendenti in ascensore?

Qui abbiamo una v.c. $\mathbf{Bin}(n = 500,\, \pi = 0.0022)$. Siamo interessati a $k \neq 0$.

$$\mathbf{P}(\mathcal{Y} \neq 0) = 1 - \mathbf{P}(\mathcal{Y} = 0) = 1 - (1 - 0.0022)^{500} = 0.33$$

Si noti che c'erano le condizioni necessarie per utilizzare l'approssimazione della v.c. Binomiale alla v.c. di Poisson. Ciò è lasciato come esercizio.

4.1 Stima dei percentili della v.c. Normale

La tavola che utilizziamo non fa altro che mettere in relazione ciascun percentile della v.c. Normale standardizzata considerato, con l'area di probabilità situata alla sinistra di tale percentile. Di conseguenza, se vogliamo calcolare il 70° percentile di una $\mathbf{N}(0, 1)$, è sufficiente prendere la tavola, cercare un valore prossimo a 0.7 e vedere a che valore di $\mathcal{Z}$ corrisponde (≈ 0.524). Per il 30° percentile, possiamo sfruttare la simmetria attorno alla media (50° percentile), quindi $\mathcal{Z}_{0.3} = -\mathcal{Z}_{0.7} \approx -0.524$.

Tuttavia, nella maggior parte dei casi, la nostra distribuzione d'interesse è Normale con una certa media μ ed una certa varianza σ^2. In questo caso, i percentili vanno determinati sulla scala originaria. Per far ciò, dobbiamo tornare dalla scala di $\mathcal{Z}$ a quella di X tramite:

1. *de–standardizzazione*:
   - se $p < 0.5 \implies \Delta_X = \mathcal{Z}_{1-p} \times \sigma$
   - se $p > 0.5 \implies \Delta_X = \mathcal{Z}_p \times \sigma$
2. *de–centratura*:
   - se $p < 0.5 \implies X_p = \mu - \Delta_X$
   - se $p > 0.5 \implies X_p = \mu + \Delta_X$

Prendiamo in considerazione la distribuzione peso del punto precedente, che ha media 72 e varianza 400. Si determinino i seguenti percentili.

$p < 0.5$	$p > 0.5$
Il 1° decile: $X_{0.1} = \overbrace{\underbrace{72}_{\mu} -}^{\text{de-centrat.}} \overbrace{\underbrace{1.28}_{\mathcal{Z}_{1-0.1}} \times \underbrace{20}_{\sigma}}^{\text{de-standard.}} = 46.4$ Kg	**Il 9° decile:** $X_{0.9} = \overbrace{72 +}^{\text{de-centrat.}} \overbrace{\underbrace{1.28}_{\mathcal{Z}_{0.9}} \times 20}^{\text{de-standard.}} = 97.6$ Kg
Il 1° percentile: $X_{0.01} = \overbrace{72 -}^{\text{de-centrat.}} \overbrace{\underbrace{2.325}_{\mathcal{Z}_{1-0.01}} \times 20}^{\text{de-standard.}} = 25.5$ Kg	**Il 99° percentile:** $X_{0.99} = \overbrace{72 +}^{\text{de-centrat.}} \overbrace{\underbrace{2.325}_{\mathcal{Z}_{0.99}} \times 20}^{\text{de-standard.}} = 138.5$ Kg
Il 1° quartile: $X_{0.25} = \overbrace{72 -}^{\text{de-centrat.}} \overbrace{\underbrace{0.675}_{\mathcal{Z}_{1-0.25}} \times 20}^{\text{de-standard.}} = 58.5$ Kg	**Il 3° quartile:** $X_{0.75} = \overbrace{72 +}^{\text{de-centrat.}} \overbrace{\underbrace{0.675}_{\mathcal{Z}_{0.75}} \times 20}^{\text{de-standard.}} = 85.5$ Kg

5 Esercizio in cui sfruttare la disuguaglianza di Čebyšëv

Abbiamo appena constatato che, considerando una distribuzione Normale di media 72 e varianza 400, il 98% centrale dei valori si colloca tra 25.5 e 138.5.

Supponiamo ora di non sapere che tale distribuzione è Normale, ma conosciamo solo la media e la varianza: **qual è la percentuale minima di valori che sono compresi tra 25.5 e 138.5?**

Per avere una stima pessimistica della probabilità in un intervallo di una qualunque distribuzione con media μ e varianza σ^2, possiamo avvalerci della disuguaglianza di Čebyšëv:

$$\mathbf{P}\left(\mu - k\sigma \leq \text{v.c.} \leq \mu + k\sigma\right) \geq 1 - \frac{1}{k^2}$$

Dobbiamo trovare k:

$$72 + k \times 20 = 138.5 \iff k = 2.325 = \mathcal{Z}_{0.99}$$

Dunque:

$$\mathbf{P}\left(25.5 \leq \text{v.c.} \leq 138.5\right) \geq 1 - \frac{1}{2.325^2} = 0.815$$

Almeno l'81.5% dei valori di una distribuzione con media 72 e varianza 400 è compreso tra 25.5 e 138.5.