

Introduzione all'inferenza

Maura Mezzetti

Inferenza

L'inferenza statistica è un insieme di metodi con cui si cerca di trarre una conclusione sulla popolazione sulla base di alcune informazioni ricavate da un campione estratto da quella popolazione.

Concetti fondamentali

Popolazione: è il collettivo di unità statistiche, individui, famiglie, gruppi e città ecc

Campione è un sottoinsieme della popolazione

La statistica

Descrittiva

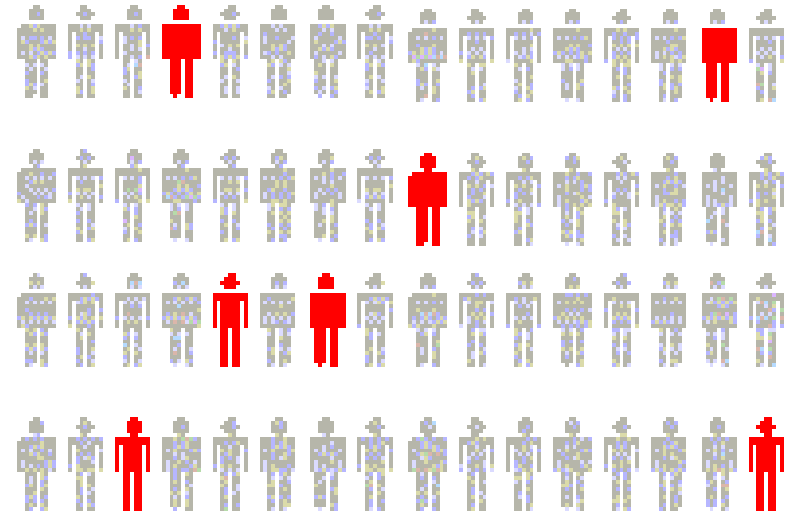
- Descrive e Sintetizza le caratteristiche principali di un collettivo di unità statistiche

Inferenza

- Come generalizzare dal campione all'intera popolazione

Statistica Inferenziale

- Stima
- Verifica di Ipotesi



Arriva a conclusioni su parametri della popolazione

Campione Casuale

Un campione casuale è un sottoinsieme
della popolazione di interesse

Con reimmissione

Senza reimmissione

Campione Casuale

Un campione casuale di ampiezza n , n variabili aleatorie $X=(X_1, X_2, \dots, X_n)$

Tra di loro identicamente distribuite e indipendenti con distribuzione marginale $f(x)$

Campione osservato

Un campione casuale osservato sarà denotato con $\mathbf{x}=(x_1, x_2, \dots, x_n)$

Lo ***spazio campionario*** (l'insieme di tutte le possibili realizzazioni) viene indicato con Ω .

Può essere discreto o continuo a seconda che X sia discreta o continua. Data l'indipendenza delle n variabili aleatorie, la distribuzione del campione risulta

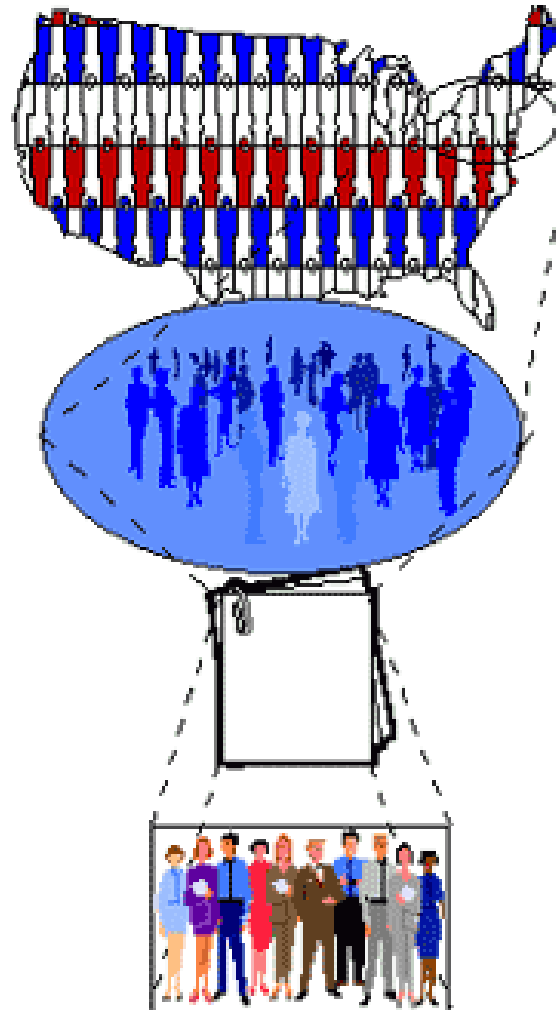
$$f(x) = f(x_1) \times f(x_2) \times \dots \times f(x_n) = \prod_{i=1}^n f(x_i)$$

Who do you want to generalize to?

What population can you get access to?

How can you get access to them?

Who is in your study?



The Theoretical Population

The Study Population

The Sampling Frame

The Sample

Fasi del processo di inferenza

- definizione del problema
- individuazione di un opportuno **modello** teorico
- estrazione del **campione**
- **raccolta e analisi** dei dati
- **generalizzazione**

Esempio

Sia X una variabile aleatoria con la seguente distribuzione:

x	0	1	2	3
$f(x)$	0.2	0.3	0.4	0.1

Supponiamo di estrarre un campione di ampiezza 2, le possibili realizzazioni sono 16

x_1	x_2	$f(x_1)$	$f(x_2)$	$f(x_1, x_2)$
0	0	0.2	0.2	0.04
0	1	0.2	0.3	0.06
0	2	0.2	0.4	0.08
0	3	0.2	0.1	0.02
....				
2	0	0.4	0.2	0.08
2	1	0.4	0.3	0.12
....	...			
3	3	0.1	0.1	0.01

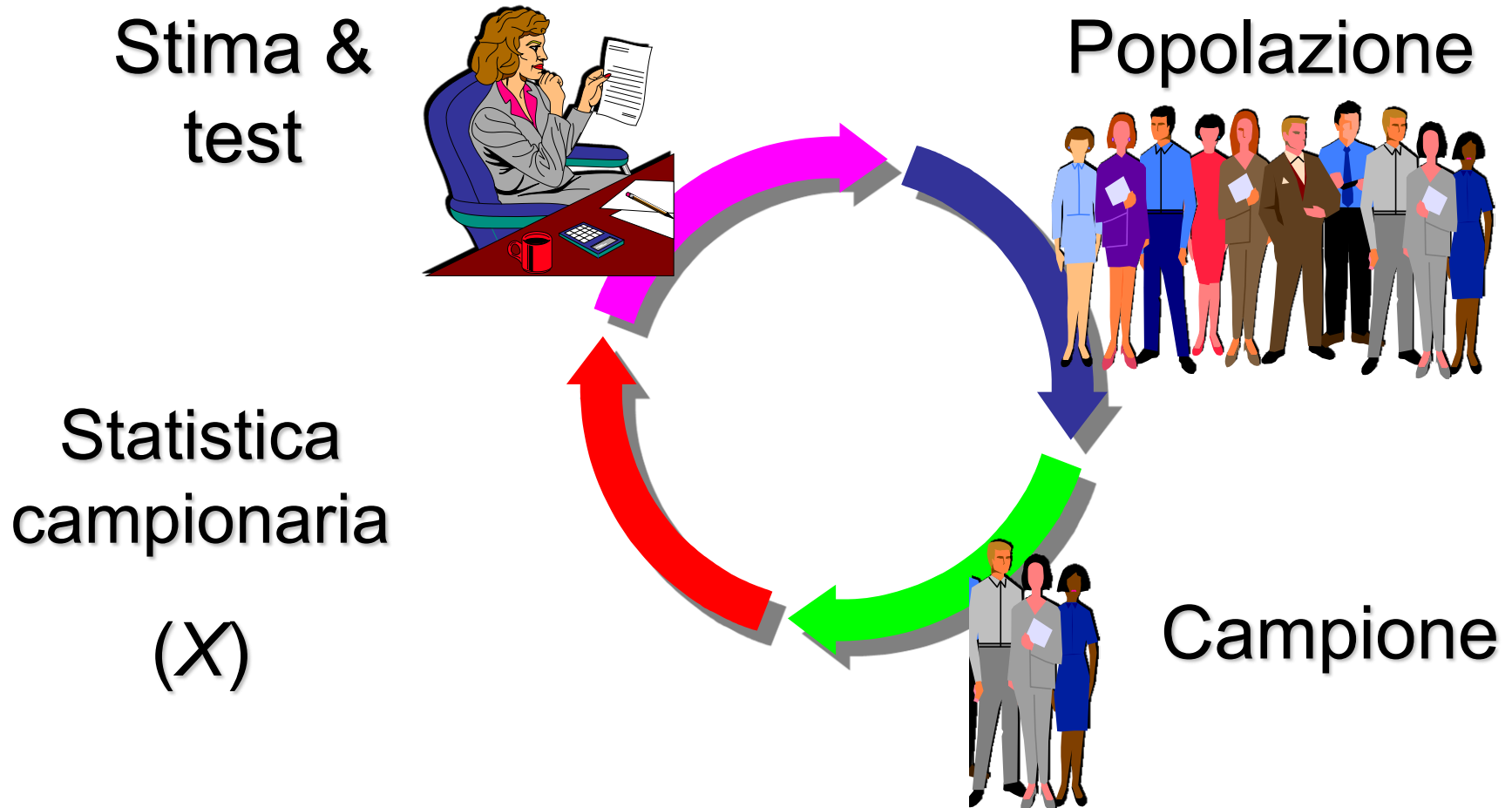
Concetti fondamentali

Parametro è una caratteristica numerica della popolazione (generalmente ignota)

Statistica è una funzione numerica del campione che NON contiene parametri ignoti

Stimatore è una statistica usata per stimare un parametro (ad esempio: media campionaria, proporzione campionaria)

Processo Inferenziale



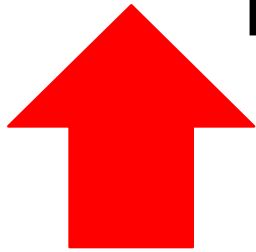
Parametro e Statistica

Un Esempio

“La media di IQ della popolazione degli studenti è 102”

Parametro

Statistica



“La media IQ di studenti di Tor Vergata è 102”

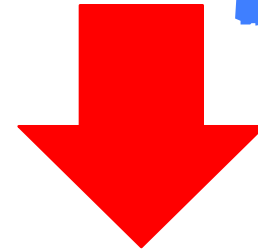
Stima



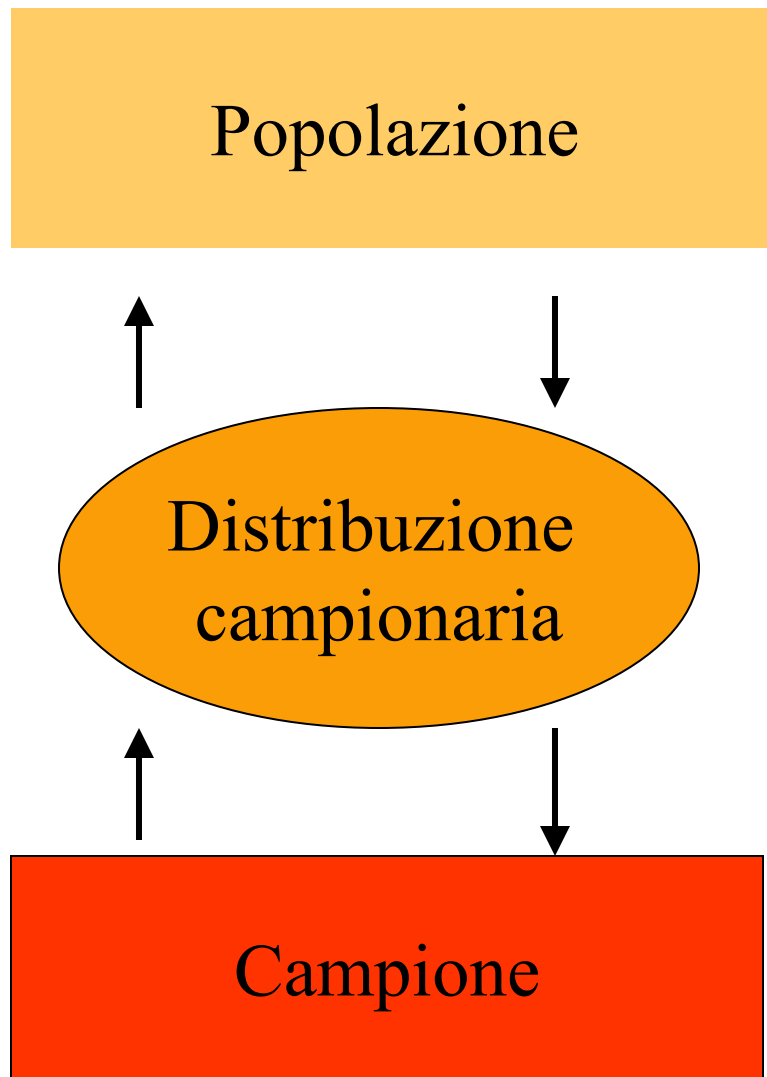
$X = \text{“IQ”}$



Popolazione



Campione ($X_1, X_2, \dots, X_n$)



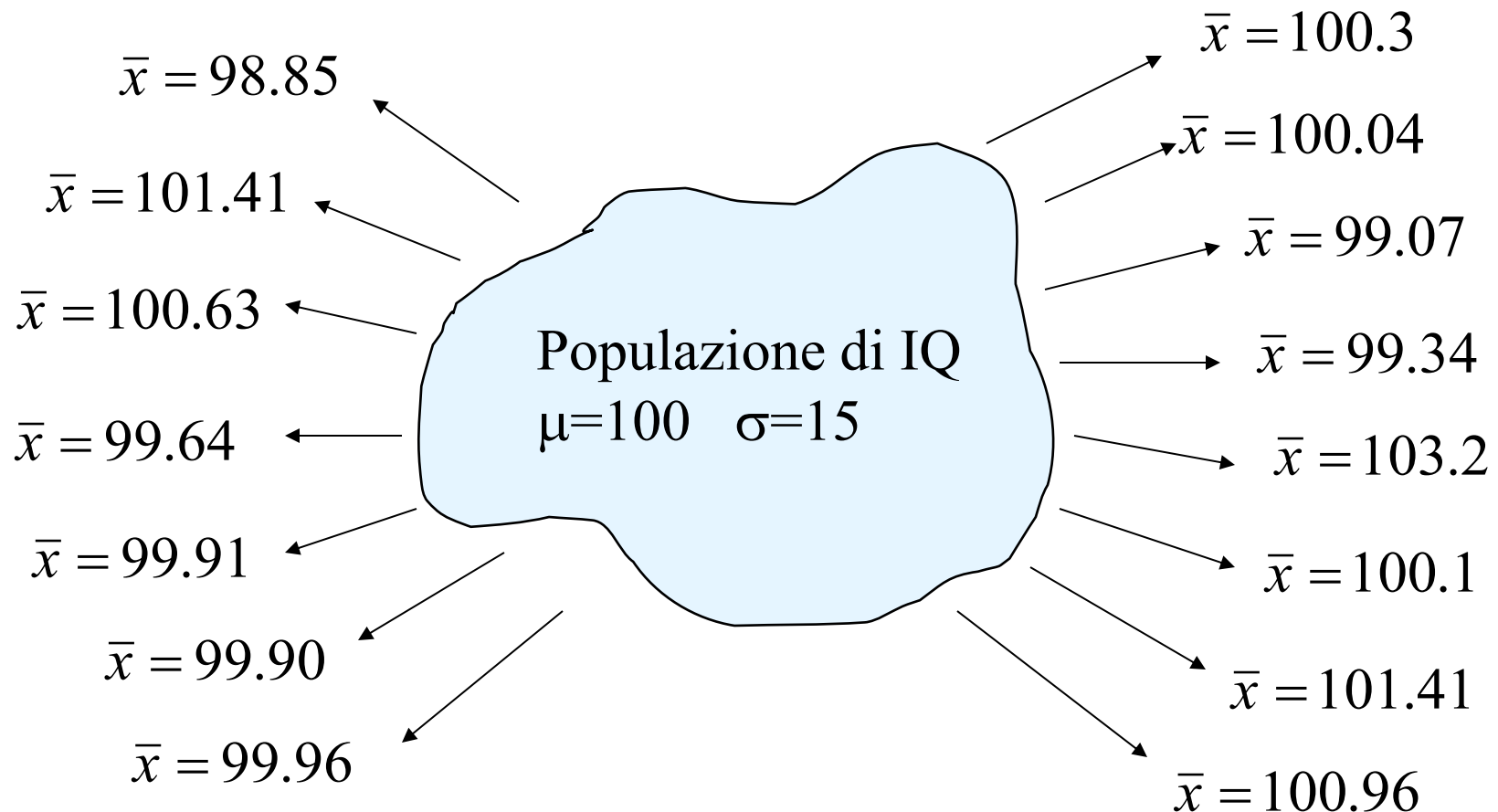
- **Distribuzione del campione:** empirica, nota
- **Distribuzione della Popolazione:** sconosciuta (scopo dell'inferenza è raccogliere informazione sulla popolazione)
- **Distribuzione campionaria** (non empirica, teorica)

x_1	x_2	$f(x_1)$	$f(x_2)$	$f(x_1, x_2)$	Media
0	0	0.2	0.2	0.04	0
0	1	0.2	0.3	0.06	0,5
0	2	0.2	0.4	0.08	1
0	3	0.2	0.1	0.02	1,5
....					
2	0	0.4	0.2	0.08	1
2	1	0.4	0.3	0.12	1,5
....	...				
3	3	0.1	0.1	0.01	3

Probabilità e statistica

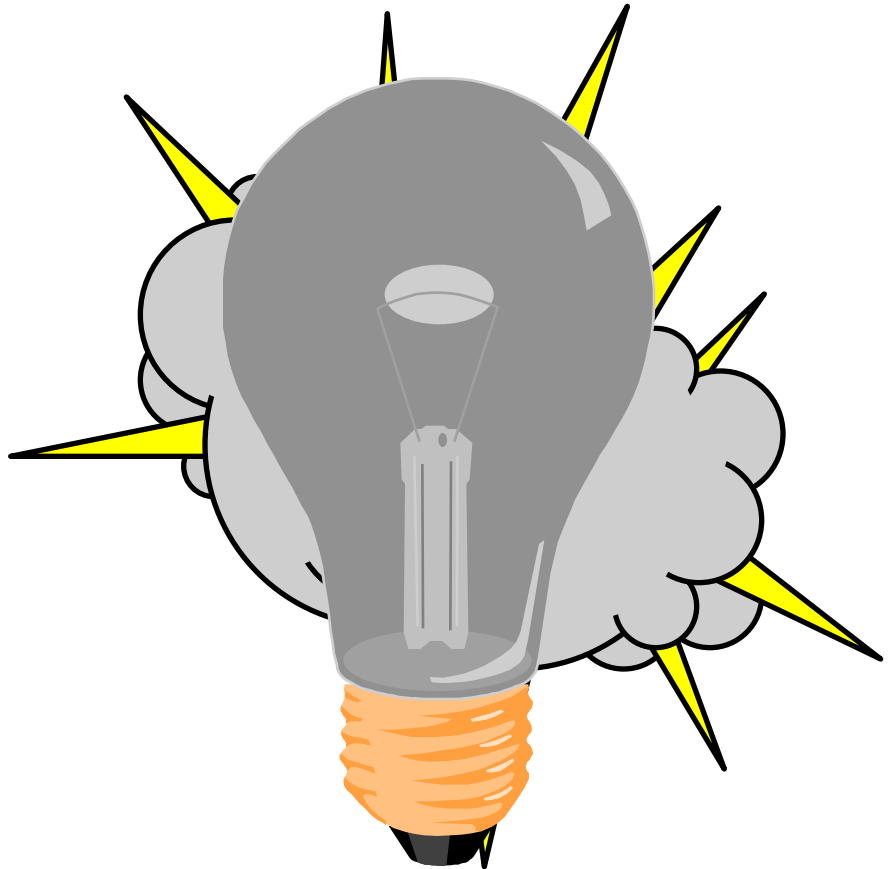
- **Probabilità** inizia dalla popolazione, descritta da una distribuzione di probabilità, e *predice* cosa succede nel campione estratto dalla popolazione.
- **Statistica** inizia da un campione di dati e descrive i dati osservati, scopo fare inferenza circa la popolazione

Supponiamo campioni di ampiezza $n=400$ estratti dalla popolazione di IQ



La media campionaria è una variabile aleatoria

1. $\bar{x}$ varia tra campione e campione
2. Dal momento che il campione è casuale, $\bar{x}$ varia con il processo di campionamento



Come si distribuisce la media campionaria?

$$\frac{\sum X_i}{n}$$

Se X è una variabile casuale Normale, anche la Somma di variabili normali è Normale:

La media campionaria si distribuisce come una Gaussiana!

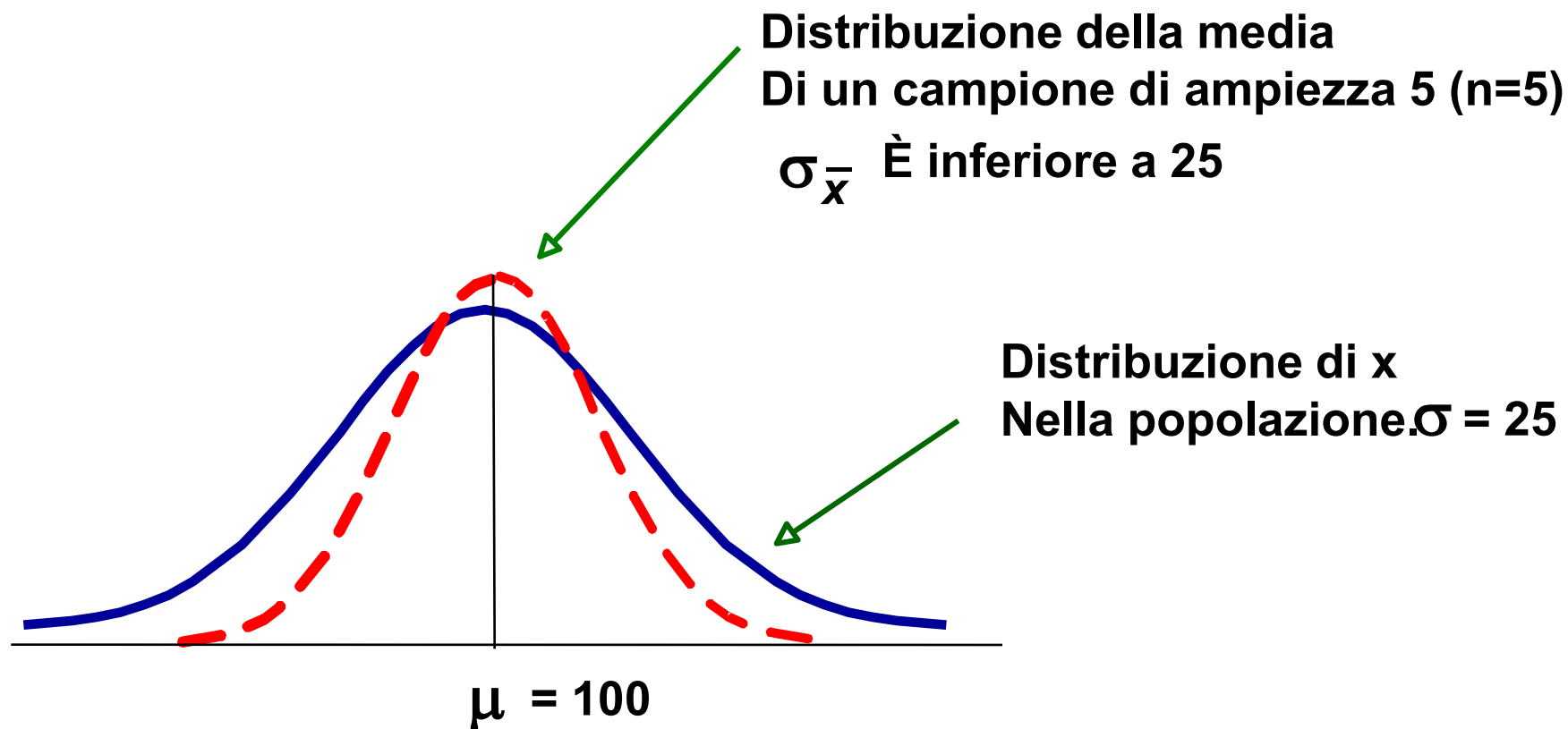
$$X \sim N(\mu, \sigma) \quad E(X) = \mu$$

$$E(\bar{X}) = E\left(\frac{\sum_{i=1}^n X_i}{n}\right) = \frac{E\left(\sum_{i=1}^n X_i\right)}{n} = \frac{\sum_{i=1}^n E(X_i)}{n} = \frac{n\mu}{n} = \mu$$

$$E(\bar{X}) = \mu$$

$$Var(\bar{X}) = Var\left(\frac{\sum_{i=1}^n X_i}{n}\right) = \frac{Var\left(\sum_{i=1}^n X_i\right)}{n^2} = \frac{\sum_{i=1}^n Var(X_i)}{n^2} = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

Distribuzione Normale e distribuzione campionaria



Distribuzione della Media Campionaria nel caso di popolazione Normale



La media della media campionaria coincide con la media della popolazione

$$\mu_{\bar{x}} = \mu$$



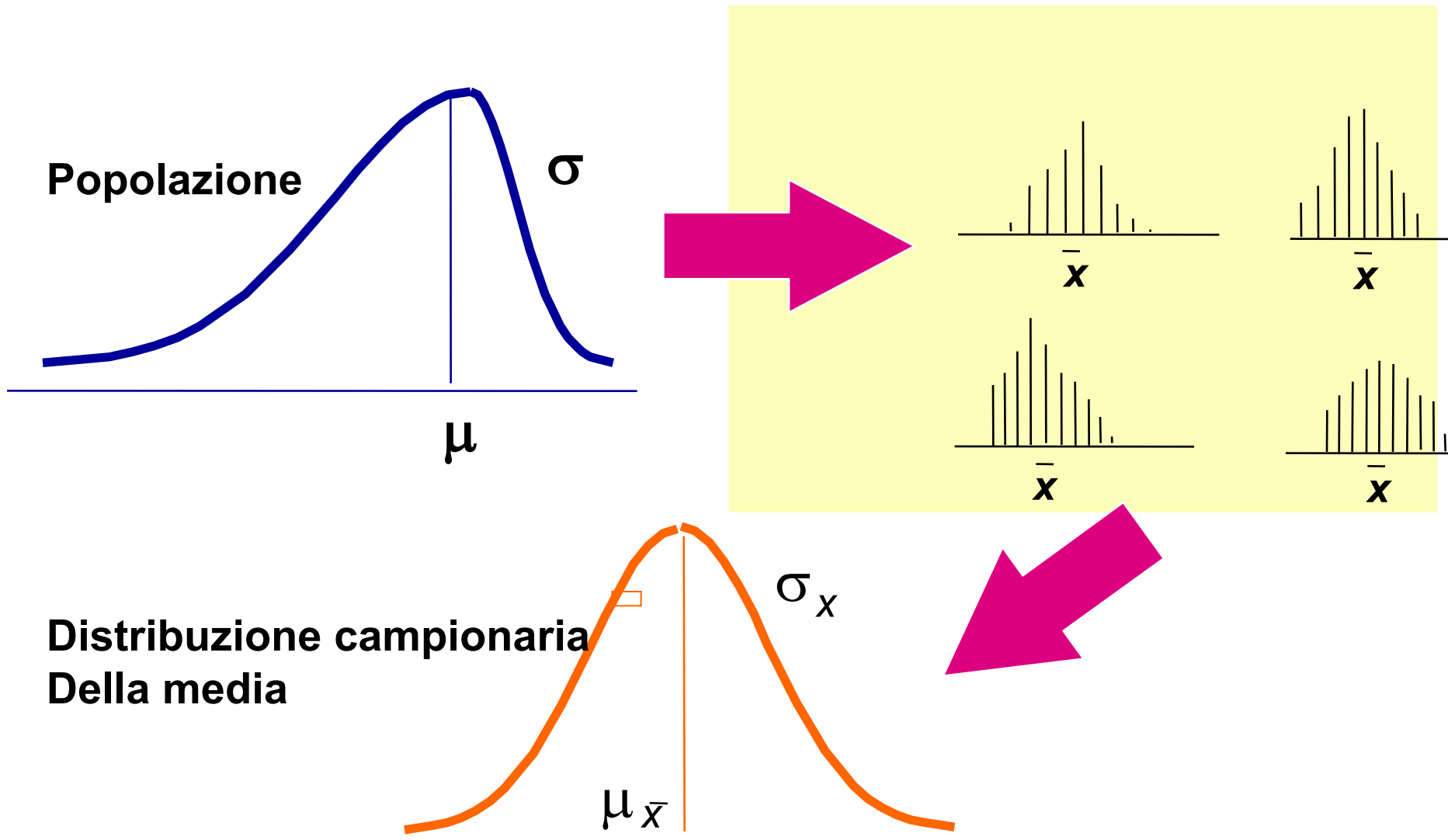
La deviazione standard del campione
Coincide con la deviazione standard della
Popolazione diviso per l'ampiezza campionaria

$$\sigma_{\bar{x}} = \frac{\sigma_X}{\sqrt{n}}$$



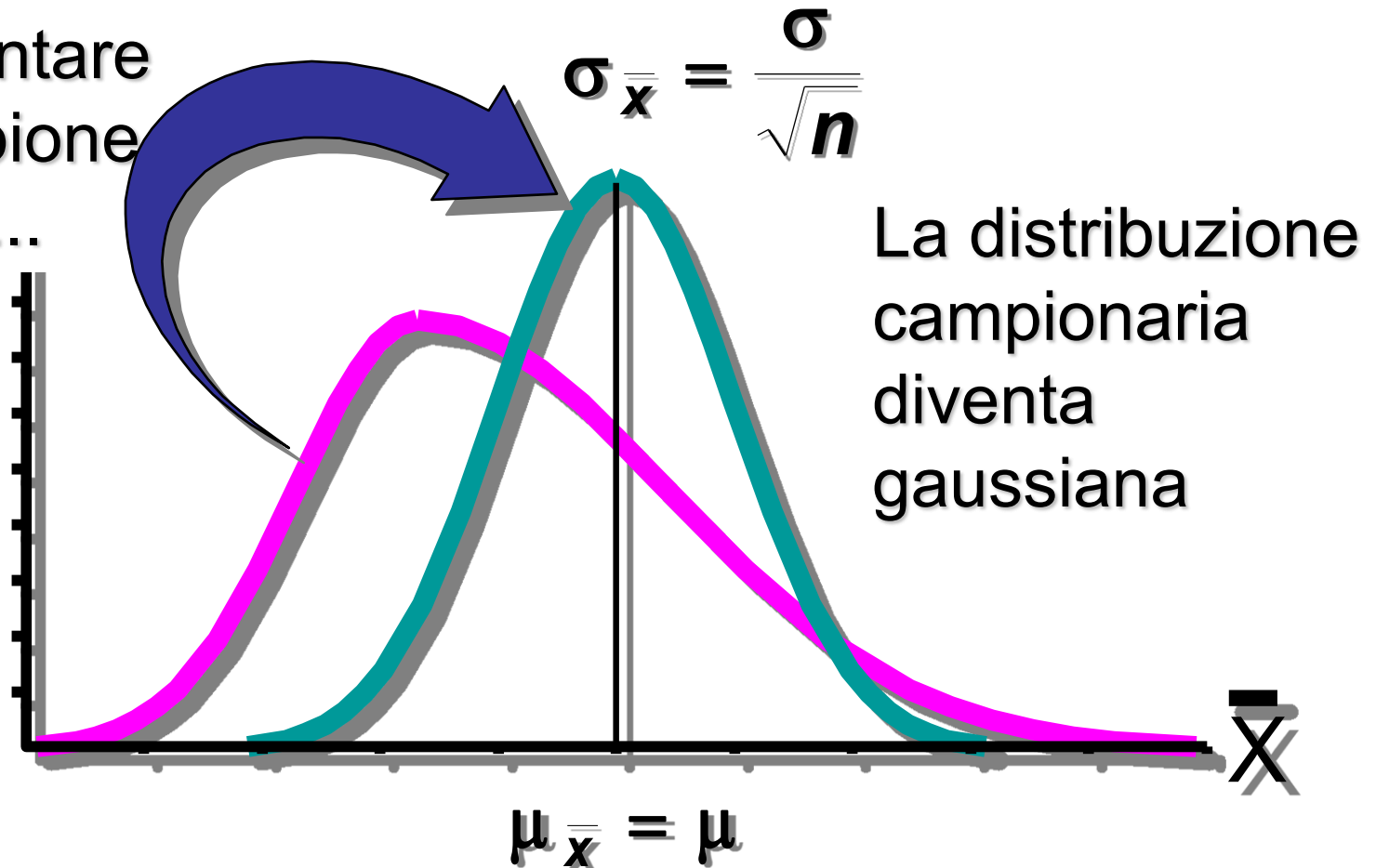
La distribuzione della media campionaria è normale

Confronto tra distribuzione popolazione E distribuzione campione



Teorema del Limite Centrale

All'aumentare
del campione
($n \geq 80$) ...



Teorema del limite centrale

Il teorema del limite centrale afferma che se n è sufficientemente elevato la distribuzione della media campionaria è approssimativamente normale, $\bar{X} \sim N(\mu, \sigma^2/n)$.

Prima di enunciare questo teorema premettiamo la nozione di convergenza in distribuzione.

Definizione: convergenza in distribuzione

Una successione di v.c. $\{Z_n, n = 1, 2, \dots\}$, con funzione di ripartizione $\{F_n(z), n = 1, 2, \dots\}$, converge in distribuzione alla v.c. $Z \sim F(z)$ (e si scrive $Z_n \rightarrow_d Z$), se, per ogni punto z in cui $F(z)$ è continua,

$$\lim_{n \rightarrow \infty} F_n(z) = F(z),$$

ovvero

$$\lim_{n \rightarrow \infty} P(Z_n \leq z) = P(Z \leq z).$$

Teorema:

Siano $\{X_1, X_2, \dots, X_n\}$ n v.c. indipendenti e identicamente distribuite (IID) con $E(X_i) = \mu < \infty$ e varianza $V(X_i) = \sigma^2 < \infty$.

Sia $\bar{X}_n = \frac{1}{n} \sum_i X_i$ la media campionaria delle n v.c.

Allora, la successione $\{Z_n, n = 1, 2, \dots\}$, con

$$Z_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} = \sqrt{n} \frac{\bar{X}_n - \mu}{\sigma}$$

converge in distribuzione ad una v.c. normale standard,

$Z_n \rightarrow_d Z \sim N(0, 1)$, ovvero,

$$\lim_{n \rightarrow \infty} P \left(\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \leq z \right) = \Phi(z).$$

Esercizio 3

Un'industria alimentare possiede un macchinario atto a produrre confezioni di pasta di peso medio 0,5 Kg e scarto quadratico medio pari a 0,005 Kg.

Si assume che la variabile aleatoria peso di una confezione si distribuisca come una normale e che i pesi delle singole confezioni siano indipendenti fra loro.

Esercizio 3

Calcolare:

- a) la probabilità che una confezione scelta a caso abbia peso compreso nell'intervallo 0,49-0,51
- b) la probabilità che il peso complessivo di 100 confezioni prese a caso sia superiore a 50,1 Kg
- c) la probabilità che una partita di merce venga rifiutata se su 10 confezioni della partita almeno 3 risultano di peso inferiore a 0,49 Kg

Soluzione 3

$$\begin{aligned} P(0.49 < X_i \leq 0.51) &= P(X_i \leq 0.51) - P(X_i \leq 0.49) = \\ &= P\left(\frac{X_i - 0.50}{0.005} \leq \frac{0.51 - 0.50}{0.005}\right) - P\left(\frac{X_i - 0.50}{0.005} \leq \frac{0.49 - 0.50}{0.005}\right) = \\ &= P(Z_i \leq 2) - P(Z_i \leq -2) = 0.9772 - 1 + 0.9772 = 0.9544 \end{aligned}$$

$$Y = \sum_{i=1}^{100} X_i \approx N(100 \times 0.50, 10 \times 0.005)$$

$$\begin{aligned} P(Y > 50.1) &= P\left(\frac{Y - 50}{0.05} > \frac{50.1 - 50}{0.05}\right) = P(Z > 2) = \\ &= 1 - 0.9772 = 0.0228 \end{aligned}$$

Esercizio 4

L'Alitalia sostiene che 5% delle persone che hanno prenotato il volo non si presenteranno al check in.

Se la compagnia ha venduto 240 biglietti per un volo che ha solo 233 posti a sedere,

Qual è la probabilità che tutti i passeggeri che si presentano abbiano un posto a sedere?

Soluzione 4

X “presentarsi/non presentarsi al check in”

$X \sim \text{Bernoulli}(0.95)$

$X_1, X_2, \dots, X_{240}$ campione i.i.d. come X

$$\hat{p} = \frac{\sum_{i=1}^{240} X_i}{240}$$

$$\hat{p} \sim N\left(0.95, \sqrt{\frac{0.95 \times (1 - 0.95)}{240}}\right)$$

$$Y = \sum_{i=1}^{240} X_i$$

$$Y \sim N\left(240 \times 0.95, \sqrt{240 \times 0.95 \times (1 - 0.95)}\right)$$

$$P(Y \leq 233) = P\left(\frac{Y - 240 \times 0.95}{3.3764} \leq \frac{233 - 228}{3.3764}\right) = P(Z \leq 1.48) = 0.93$$

Esercizio 3

Un'industria alimentare possiede un macchinario atto a produrre confezioni di pasta di peso medio $0,5 \text{ Kg}$ e scarto quadratico medio pari a $0,005 \text{ Kg}$.

Si assume che i pesi delle singole confezioni siano indipendenti fra loro.

Qual è la probabilità che il peso complessivo di 100 confezioni prese a caso sia superiore a $50,1 \text{ Kg}$

Soluzione 3

$$\begin{aligned}P(0.49 < X_i \leq 0.51) &= P(X_i \leq 0.51) - P(X_i \leq 0.49) = \\&= P\left(\frac{X_i - 0.50}{0.005} \leq \frac{0.51 - 0.50}{0.005}\right) - P\left(\frac{X_i - 0.50}{0.005} \leq \frac{0.49 - 0.50}{0.005}\right) = \\&= P(Z_i \leq 2) - P(Z_i \leq -2) = 0.9772 - 1 + 0.9772 = 0.9544\end{aligned}$$

$$Y = \sum_{i=1}^{100} X_i \approx N(100 \times 0.50, 10 \times 0.005)$$

$$\begin{aligned}P(Y > 50.1) &= P\left(\frac{Y - 50}{0.05} > \frac{50.1 - 50}{0.05}\right) = P(Z > 2) = \\&= 1 - 0.9772 = 0.0228\end{aligned}$$

Esercizio

Un test attitudinale rivolto agli studenti della scuola superiore è costruito in modo tale che i punteggi del test abbiano media $\mu=90$ e $\sigma=20$. Gli studenti in una scuola sono assegnati casualmente a varie classi di uno stesso corso. In una di queste classi composta da 100 studenti, se l'assegnazione degli studenti è casuale, qual è la probabilità di ottenere nel test una media minore o uguale a 86?

Esercizio

La media campionaria avrà una distribuzione approssimativamente normale, con media 90 e $\sigma = 2$.

Si ottiene:

$$P(\bar{X} \leq 86) = P\left(Z \leq \frac{86 - 90}{2}\right) = P(Z \leq -2) = 1 - P(Z \leq 2) = 0.0228$$

Esercizio

Siano $X_1, \dots, X_{144}$ variabili aleatorie iid con legge bernoulliana di parametro 0.4. Se

$$Y = \sum_{i=1}^{144} X_i$$

si fornisca una valutazione approssimata di $P(Y \leq 60)$.

Esercizio

$$Y = \sum_{i=1}^{144} X_i$$

si fornisca una valutazione approssimata di $P(Y \leq 60)$.

$$P(Y \leq 60) = P\left(\frac{\sum_{i=1}^{144} X_i - 144 \times 0.4}{\sqrt{144 \times 0.4 \times 0.6}} \leq \frac{60 - 144 \times 0.4}{\sqrt{144 \times 0.4 \times 0.6}}\right) =$$

$$P(Z \leq 0.408) = 0.6584$$

Esercizio

Una ditta di trasporti internazionali possiede 100 tir dello stesso tipo. Ogni tir percorre una media di 400 km al giorno con una deviazione standard di 40 km. Supponendo che i giorni lavorativi in un anno siano 312,

a) quanti chilometri percorre mediamente un tir in un anno?

- Supponendo che la durata dei pneumatici da tir sia in media 125000 km,

b) qual è la probabilità che un tir debba cambiare i pneumatici dopo un anno?

a) quanti chilometri percorre mediamente un tir in un anno?

$$400 \times 312 = 124800$$

Supponendo che la durata dei pneumatici da tir sia in media 125000 km,

b) qual è la probabilità che un tir debba cambiare i pneumatici dopo un anno?

$$P\left(\frac{\sum X - 124800}{\sqrt{312 \times 40^2}} \geq \frac{125000 - 124800}{\sqrt{312 \times 40^2}}\right) = P(Z \geq 0.28) = 0.3897$$