Università degli Studi di Palermo
Dipartimento di Ingegneria Informatica

C.I. 1 – "Informatica ed Elementi di Statistica"
2 c.f.u.

Anno Accademico 2009/2010

Docente: ing. Salvatore Sorce

Elementi di Statistica

Università degli Studi di Palermo

Statistica - definizioni

- Ramo del sapere che impiega strumenti logici e matematici per la raccolta, il raggruppamento e l'interpretazione dei dati

- Scienza che ha come fine lo studio quantitativo e qualitativo di un "collettivo". Studia i modi (descritti attraverso formule matematiche) in cui una realtà fenomenica - limitatamente ai fenomeni collettivi - può essere sintetizzata e quindi compresa

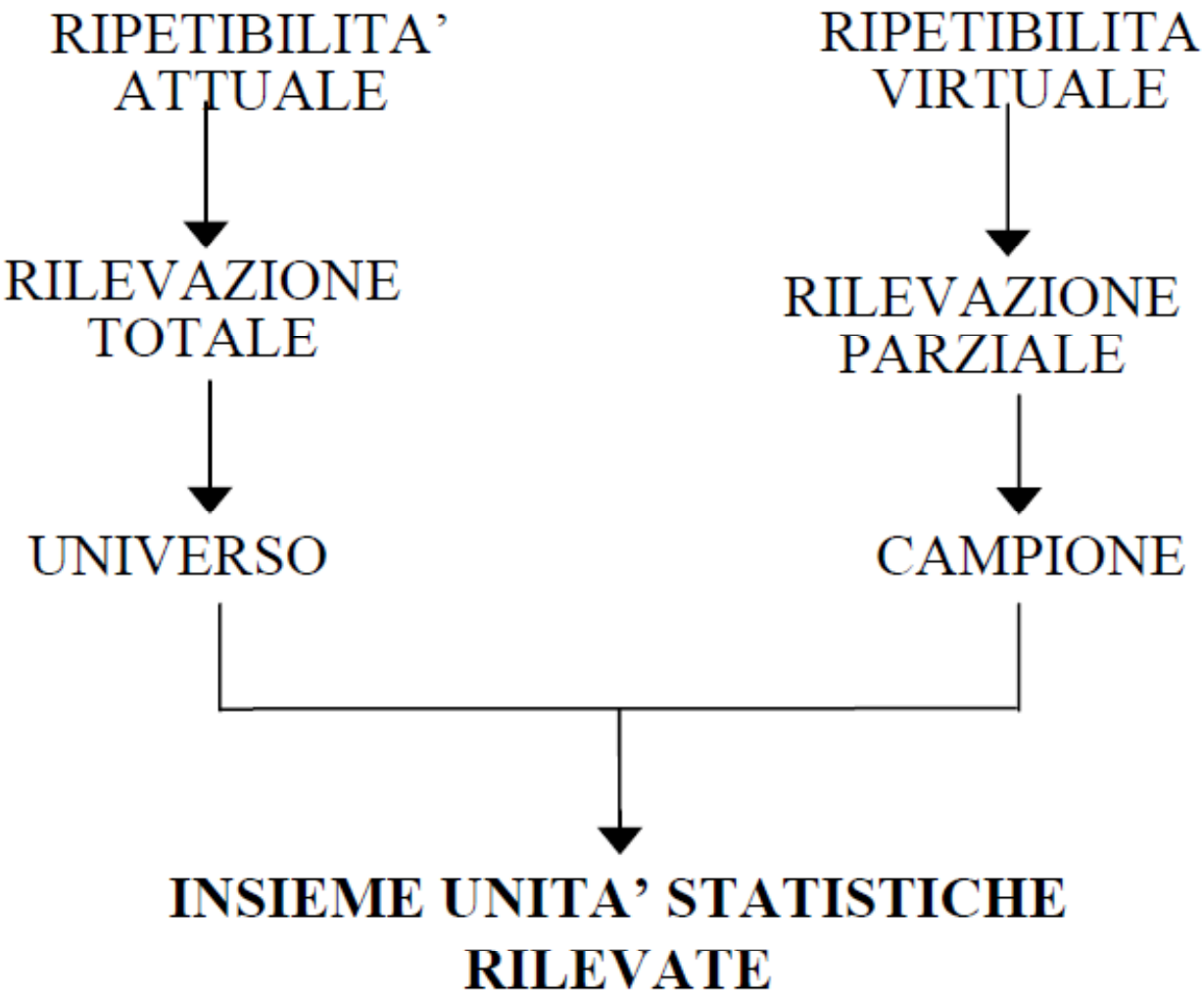
Statistica

Si occupa di fenomeni ripetibili del mondo reale che si manifestano con determinazioni non costanti (presenza di variabilità)

Si distingue tra:

1. Ripetibilità *attuale*: tutte le manifestazioni di interesse del fenomeno si sono già realizzate (es. fenomeni demografici)
2. Ripetibilità *virtuale*: non tutte le manifestazioni si sono realizzate (es. unità prodotte da un dato processo produttivo)

Statistica


RIPETIBILITA' ATTUALE
RIPETIBILITA' VIRTUALE
RILEVAZIONE TOTALE
RILEVAZIONE PARZIALE
UNIVERSO
CAMPIONE
INSIEME UNITA' STATISTICHE RILEVATE

Rilevazione

1. Individuazione di uno o più CARATTERI sui quali acquisire le informazioni

2. Individuazione delle UNITA' STATISTICHE portatori del carattere in studio

3. Procedimento di misurazione del carattere che porta alla individuazione delle MODALITA' con cui il carattere si presenta

Modalità di rilevazione

- Fenomeni QUALITATIVI
  - *si identificano tramite attributi*
- Fenomeni QUANTITATIVI
  - *si identificano tramite numeri*

Fenomeni qualitativi

- Scale nominali (o sconnesse o categoriali):
  - le modalità non sono suscettibili di alcun tipo di ordinamento
- Scale ordinali (o rettilinee):
  - le modalità presentano in via naturale un ordine

Fenomeni quantitativi

- Discreti:
  - caratteri numerabili, modalità ottenibile tramite un'operazione di conteggio (classe dei numeri naturali)
- Continui:
  - caratteri misurabili, modalità ottenuta tramite un'operazione di misurazione (classe dei numeri reali)

Esempio

Qualitativa nominale

Qualitativa ordinale

ID	CORSO LAUREA	SESSO	MEDIA VOTI	CREDITI	RENDIMENTO
1	SAM	M	22	6	discreto
2	SAM	F	24	71	buono
3	SAM	M	21	19	discreto
4	SAM	F	26	27	buono
5	SAM	F	27	9	ottimo
6	SAM	M	26	10	buono
7	SAM	F	25	18	buono
8	SAM	M	24	27	buono
9	SAM	F	27	10	ottimo
10	SAM	F	24	17	buono
11	SAM	M	26	18	buono
12	SAM	M	30	18	ottimo
13	SAM	F	29	84	ottimo
14	SPO	M	27	27	ottimo
15	SPO	F	23	9	discreto
16	SPO	F	27	30	ottimo
17	SPO	M	28	33	ottimo
18	SPO	M	29	30	ottimo
19	SPO	F	28	48	ottimo
20	ORU	F	26	66	buono

Quantitativa continua (è una media!)

Quantitativa discreta (deriva da un conteggio)

Università degli Studi di Palermo

Distribuzioni di frequenza

Sintesi tabellare dei caratteri statistici:

Si hanno *n dati* relativi ad un indagine condotta su *n individui;* ad ogni modalità x_i del carattere X si associa il numero di volte n_i in cui tale modalità si manifesta

n = numero delle unità statistiche rilevate

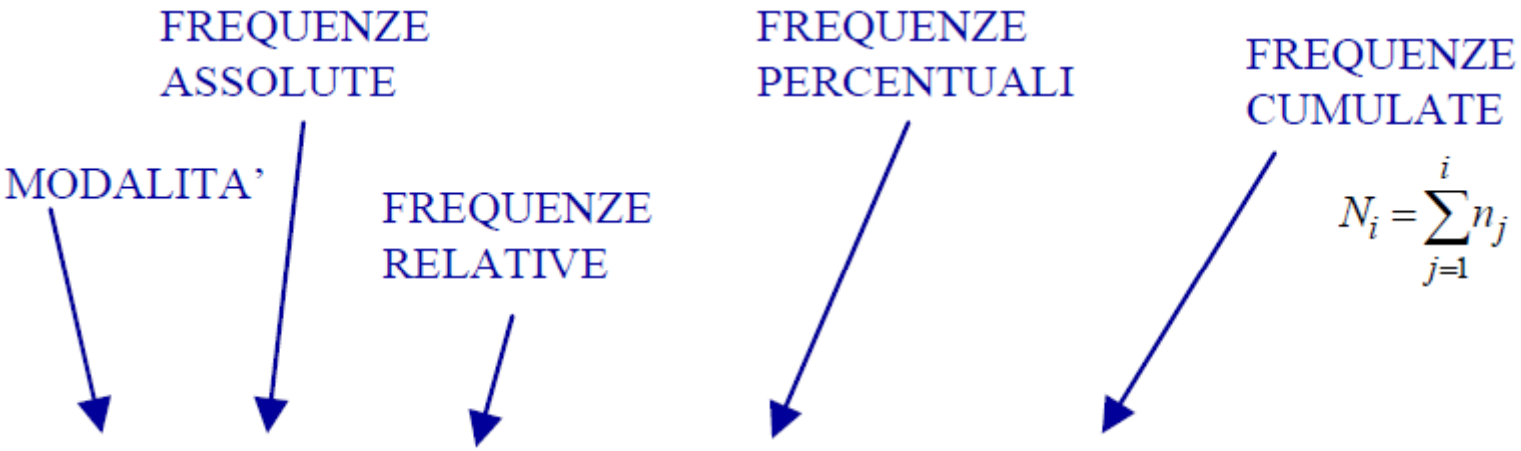
X = carattere oggetto di studio

K = num totale dei diversi valori assunti dalla variabile X (modalità)

x_i = modalità i-esima del carattere X i=1,….,k

n_i = frequenze assolute

Distribuzioni di frequenza



x_i	n_i	f_i	p_i	N_i
x_1	n_1	$n_1/n=f_1$	f_1*100	n_1
x_2	n_2	$n_2/n=f_2$	f_2*100	$n_{1+}n_2$
x_3	n_3	$n_3/n=f_3$	f_3*100	$n_{1+}n_{2+}n_3=n$
	n	1	100	

$$\begin{cases} N_1 = n_1 \\ N_k = n \\ N_i - N_{i-1} = n_i \end{cases}$$

Rappresentazioni grafiche

- Caratteri qualitativi sconnessi e rettilinei:
  - Rappresentazione tramite rettangoli
  - Grafici a torta o a settori circolari
  - Grafici a pila
- Caratteri quantitativi discreti
  - Rappresentazione tramite segmenti o bastoncini
- Caratteri quantitativi continui
  - Istogramma (o canne d'organo)
  - Poligoni di frequenza

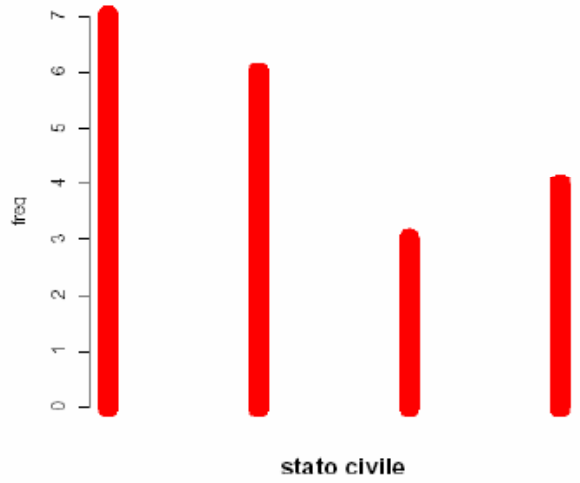
Rappresentazioni grafiche

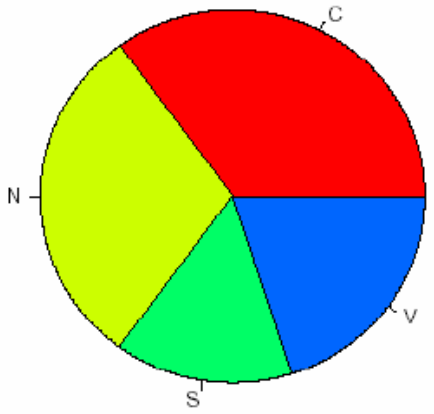
u	X	Y	Z	W
unità stat.	stato civile	livello di studi	numero di figli	peso in Kg
1	N	L	0	72.50
2	S	O	1	54.28
3	V	A	3	50.02
4	V	O	4	88.88
5	C	L	1	62.30
6	N	S	1	45.21
7	C	S	0	57.50
8	C	O	2	78.40
9	V	L	3	75.13
10	N	O	0	58.00
11	N	S	1	53.70
12	N	A	0	91.29
13	S	S	1	74.70
14	C	S	4	41.22
15	N	S	3	65.20
16	C	L	0	63.58
17	V	O	2	48.27
18	S	O	2	52.52
19	C	S	4	69.50
20	C	S	4	85.98

Rappresentazioni grafiche

Stato civile

x_i	n_i	f_i	p_i
N	6	0.30	30
C	7	0.35	35
V	4	0.20	20
S	3	0.15	15
	$n = 20$	1.00	100

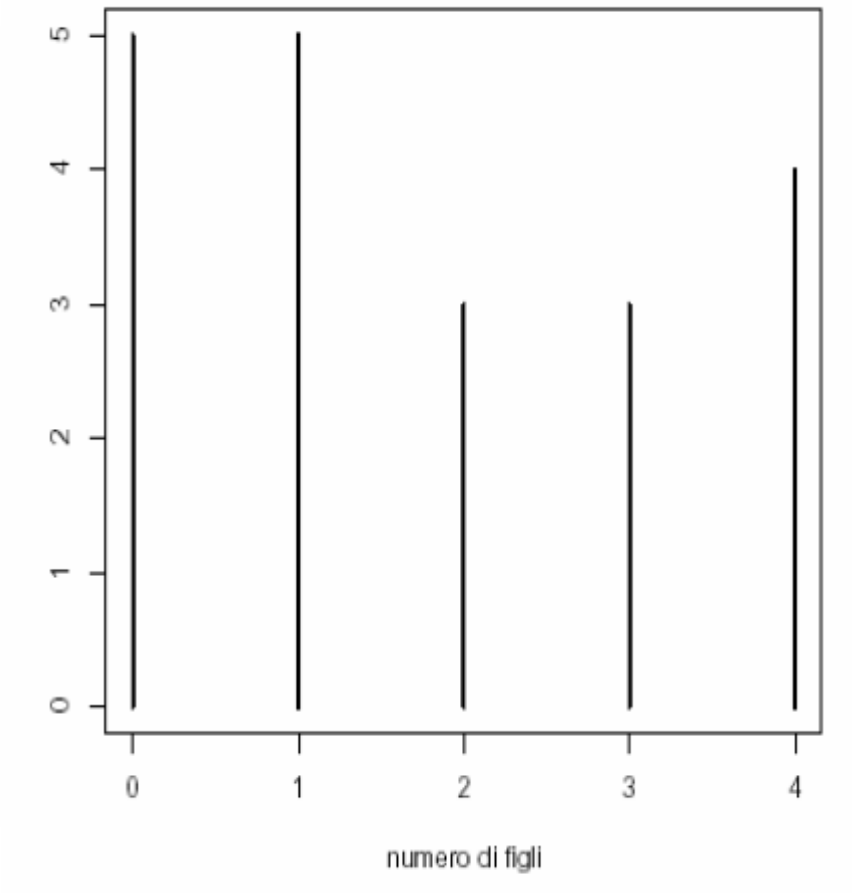




Rappresentazioni grafiche

Numero di figli

z_i	n_i	f_i	p_i	$F(x_i)$
0	5	0.25	25	0.25
1	5	0.25	25	0.50
2	3	0.15	15	0.65
3	3	0.15	15	0.80
4	4	0.20	20	1.00
	20	1.00	100	



Università degli Studi di Palermo

Indici di posizione

- Gli indici si posizione sono misure sintetiche ("valori caratteristici") che descrivono *la tendenza centrale* di un fenomeno

- *La tendenza centrale* è, in prima approssimazione, la modalità della variabile verso la quale i casi tendono a gravitare, ossia il 'baricentro' della distribuzione

Università degli Studi di Palermo

Moda o norma

- È la modalità della variabile alla quale è associata la maggior frequenza, cioè quella che si è manifestata più volte in sede di rilevazione

$$Mo = \{x_i : \max_i(n_i) \quad i = 1, \ldots, k\}$$

- Può essere calcolata per qualsiasi tipo di variabile
- È un indice elementare e non molto 'informativo'
- Una distribuzione è *unimodale* se ammette un solo valore modale, è *bimodale* se ne ammette due (ossia: se esistono due valori che compaiono entrambi con la frequenza massima nella data distribuzione), *trimodale* se ne ha tre, ecc.

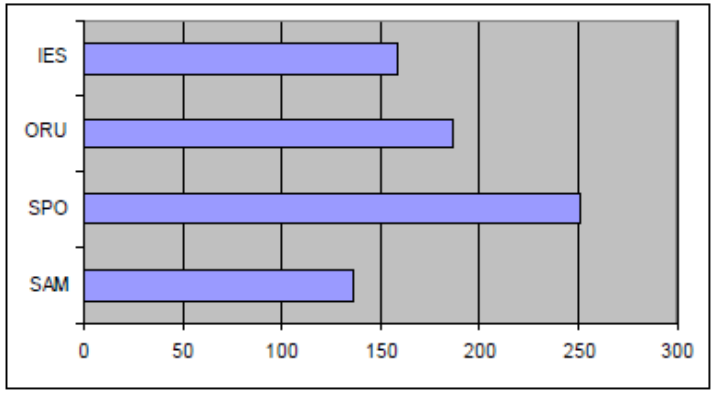
Moda

ESEMPI

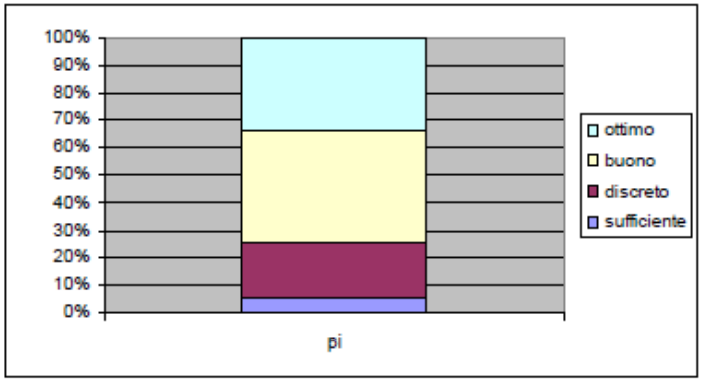
VARIABILI QUALITATIVE
SCONNESSE O RETTILINEE

CORSO	ni	fi	pi
SAM	137	0,19	19%
SPO	251	0,34	34%
ORU	186	0,25	25%
IES	159	0,22	22%
	733	1,00	100%

Mo=SPO



rendim	ni	fi	pi	Ni	Fi	Pi
sufficiente	40	0.05	5.47%	40	0.05	5.47%
discreto	150	0.21	20.52%	190	0.26	25.99%
buono	293	0.40	40.08%	483	0.66	66.07%
ottimo	248	0.34	33.93%	731	1.00	100.00%
	731	1.00	100.00%			



Mo=Buono

Università degli Studi di Palermo

Mediana

- La mediana di una variabile è la modalità (valore) che occupa la posizione centrale nella distribuzione *ordinata* della variabile.
- non può essere calcolata per le variabili sconnesse perché non posseggono in via naturale un ordine
- è un indice più informativo della moda

- Per calcolare la mediana di *n* dati:
  - si ordinano gli *n* di dati in ordine crescente o decrescente;
  - se il numero di dati è dispari la mediana corrisponde al valore centrale, ovvero al valore che occupa la posizione $(n + 1) / 2$.
  - se il numero *n* di dati è pari, la mediana è stimata utilizzando i due valori che occupano le posizione $(n / 2)$ e $(n / 2 + 1)$ (generalmente si sceglie la loro media aritmetica).

Media aritmetica

- La media è il valore caratteristico più noto e più impiegato fra quelli che rilevano la tendenza centrale
- Viene calcolata sommando i diversi valori a disposizione, i quali vengono divisi con il numero complessivo di valori.
- Viene usata per riassumere con un solo numero un insieme di dati su un fenomeno misurabile (per esempio, l'altezza media di una popolazione).

$$m_g = \frac{x_1 + x_2 + x_3 + \ldots + x_N}{N}$$

- **Può essere calcolata solo per variabili quantitative**
- ATTENZIONE: Molto spesso è comodo associare alle modalità qualitative codici numerici (es. numero di matricola, codice identificativo cliente). Nonostante la ricodifica, la variabile rimane connotata secondo la caratteristica intrinseca del fenomeno di cui essa è rilevazione.
- NON HA SENSO FARE LA MEDIA DEL NUMERO DI MATRICOLA!

Università degli Studi di Palermo

Media aritmetica ponderata (o pesata)

- Nella media aritmetica ponderata (media pesata), i singoli valori, prima di essere sommati vengono moltiplicati con il *peso* (ponderazione) a loro assegnato.
- Il peso di ciascun valore è in genere rappresentato dal numero di volte in cui i valori figurano (frequenza), ma può significare anche l'importanza (oggettiva o soggettiva) che il singolo valore riveste nella distribuzione. La divisione di conseguenza non viene fatta con il numero di valori, ma con la somma dei *pesi*.

$$M_{a,pond} = \frac{\sum_i x_i \cdot f_i}{\sum_i f_i}$$

Università degli Studi di Palermo

Percentile

- Il percentile è il valore di una variabile (aleatoria) sotto il quale si verifica una certa percentuale dell'osservazione. Ad esempio il 10° percentile è il valore sotto al quale si trovano il 10% delle osservazioni.
- In ambito discreto, valutare un percentile significa determinare il termine di un insieme ordinato tale per cui lui e tutti gli altri termini ad esso inferiori sono in quantità pari al valore di percentile visto come percentuale.
- Considerando un campione di *n* dati, ordinati in modo crescente, l'indice del k-esimo percentile è ottenuto dalla formula:

$$I_k = \lfloor 0.5 + (n * k/100) \rfloor$$

Percentile - esempio

- Calcolo del 45esimo percentile dall'insieme ordinato A={0,4,5,12,56,66,70,90,92,94,106,129,140,141,190,299,304,509,606,720,841,1022,4890,12673}
- In questo caso, *n* = 24 (numero di dati dell'insieme ordinato)

- L'indice del 45esimo percentile sarà quindi dato da:

$$I_k = \lfloor 0.5 + (n * k/100) \rfloor$$

$$I_k = \lfloor 0.5 + (24 * 45/100) \rfloor = \lfloor 11.3 \rfloor = 11$$

- quindi 106 (l'undicesimo dato dell'insieme) è il percentile cercato. Ciò esprime il fatto che il 45% dei numeri dell'insieme ha valore minore o uguale a 106.

Università degli Studi di Palermo

Indici di dispersione

- Un indice di dispersione (o indicatore di dispersione o indice di variabilità o indice di variazione) serve per descrivere sinteticamente una distribuzione statistica quantitativa, e in modo particolare la misura con la quale i suoi valori sono distanti da un valore centrale (identificato con un indice di posizione, solitamente media o mediana).

- Varianza

- Deviazione standard

Università degli Studi di Palermo

Varianza

- La varianza, detta anche media degli scarti al quadrato, viene solitamente indicata con σ^2 (dove σ è la deviazione standard).
- L'espressione della varianza, nell'ambito della statistica descrittiva, è:

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

- dove μ rappresenta la media aritmetica dei valori x_i
- La varianza è un indicatore di dispersione in quanto è nulla solo nei casi in cui tutti i valori sono uguali tra di loro (e pertanto uguali alla loro media) e cresce con il crescere delle differenze reciproche dei valori.
- Trattandosi di una somma di valori (anche negativi) al quadrato, la varianza non sarà mai negativa.

Deviazione standard

- La deviazione standard o scarto quadratico medio è un indice di derivato direttamente dalla varianza, ha la stessa unità di misura dei valori osservati (mentre la varianza ha come unità di misura il quadrato dell'unità di misura dei valori di riferimento).
- La deviazione standard misura la dispersione dei dati intorno al valore atteso (valore medio).

$$\sigma_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n}}$$