SINTESI DELLE DISTRIBUZIONI UNIVARIATE
INDICI DI VARIABILITA'

Rappresentazione dei dati { **Tabellare**, **Grafica**, **Sintetica (indici)**, **Analitica** }

Sostituire la distribuzione (successione) con un unico valore allo scopo di mettere in evidenza un particolare aspetto della distribuzione stessa

Indici di { **Posizione**, **Variabilità**, **Forma** }

CARATTERI QUANTITATIVI

Attitudine del carattere ad assumere valori differenti tra le diverse unità statistiche.

Si deve costruire un indice che

- **sia non negativo**
- **sia 0 se tutte le modalità osservate sono uguali**
- **tenda a crescere quanto più le modalità sono differenti tra loro**

(*la risposta della statistica a Trilussa*)

Se le modalità sono tutte uguali la variabilità è nulla e la statistica si disinteressa di quel carattere

INDICI DI VARIABILITA' ASSOLUTI

Si abbia una successione $(x_1, x_2, \ldots, x_j, \ldots, x_N)$ di modalità di un carattere quantitativo discreto X osservato su N unità statistiche

Es: numero di addetti rilevati in cinque imprese

Valori osservati: 9 5 4 6 1

Indici basati sul confronto tra modalità

Campo di variazione $C = x_{max} - x_{min}$

Differenza tra la modalità più grande e quella più piccola

$x_{max} = 9$ $x_{min} = 1$ C=9-1=8 addetti

Scarto interquartile $SIQ = Q_3 - Q_1$

Differenza tra la le due modalità che occupano rispettivamente il 75-esimo e il 25-esimo percentile

1 (4) 5 (6) 9

$Q_3 = 6$ $Q_1 = 4$ SIQ=6-4=2 addetti

Differenza Media Semplice

$$\Delta = \frac{\sum_i \sum_j |x_i - x_j|}{N(N-1)}$$

Sintesi (media) di tutte le possibili differenze tra i valori

	1	4	5	6	9
1	0	3	4	5	8
4	3	0	1	2	5
5	4	1	0	1	4
6	5	2	1	0	3
9	8	5	4	3	0
					72

Δ = 72÷(5×4) =72÷20 = 3,6 addetti

Mediamente ogni unità differisce dalle altre di 3,6 addetti

Indici basati su differenze rispetto a un valor medio

Devianza $SSQ = \sum_j (x_j - \bar{x})^2$

media: $\bar{x} = (1+4+5+6+9) : 5 = 25 : 5 = 5$

X	(x-5)	(x-5)^2
1	-4	16
4	-1	1
5	0	0
6	1	1
9	4	16
		34

addetti al quadrato

Varianza $\sigma^2 = \frac{\sum_i (x_j - \bar{x})^2}{N}$

$\sigma^2 = 34 : 5 = 6,8$ ***addetti al quadrato***

Scarto quadratico medio $\sigma = \sqrt{\frac{\sum_i (x_j - \bar{x})^2}{N}}$

$\sigma = \sqrt{6,8} = 2,61$

Le unità, in media, differiscono dalla media aritmetica di 2,61 addetti

PROPRIETA' DELLO SCOSTAMENTO QUADRATICO MEDIO

È l'indice che si affianca naturalmente alla media aritmetica (esprime la distanza della media aritmetica da tutti i valori e, quindi, dice quanto "bene" la media rappresenta la distribuzione)

<u>Proprietà formali</u> - Se i valori sono trasformati linearmene $y_j = \alpha + \beta\, x_j$

$\sigma_y = \beta\, \sigma_x$ ***Omogeneo - Non traslativo***

<u>Scomponibilità della varianza</u>

$x_1, x_2, \ldots, x_j,$ $x_{j+1}, x_{j+2}, \ldots, x_N$

gruppo A **gruppo B**

$\sigma^2(X) =$ ***Varianza totale =*** ***Varianza entro gruppi +***

$$\left[\sigma^2(X \in A) \times j + \sigma^2(X \in B) \times (N-j)\right] / N \quad +$$

$$\left[(M(X \in A) - M(X))^2 \times j + (M(X \in B) - M(X))^2 \times (N-j)\right] / N$$

Varianza tra gruppi

Esempio numerico

Resa per ha	*Trattamento 1*	*Trattamento 2*	*Trattamento 3*	*totale*
23	-	-	1	1
24	-	1	2	3
25	-	2	1	3
26	1	1	-	2
28	1	-	-	1
totale	2	4	4	10

$M(X) = (23\times1 + 24\times3 + 25\times3 + 26\times2 + 28\times1)/10$
$= 250/10 = 25$

$Var(X) = [(23-25)^2\times1 + (24-25)^2\times3 + (25-25)^2\times3 + (26-25)^2\times2 + (28-25)^2\times1]/10$
$= [4 + 3 + 0 + 2 + 9]/10 = 18/10 = 1{,}8$

$M(X\in T1)=27 \quad Var(X\in T1)=2/2 = 1$
$M(X\in T2)=25 \quad Var(X\in T2)=2/4 = 0{,}5$
$M(X\in T3)=24 \quad Var(X\in T3)=2/10=0{,}5$

=

VAR entro Tr $=[1\times2+0{,}5\times4+0{,}5\times4]/10=6/10=0{,}6$

+

VAR TRA Tr $= [(27-25)^2\times2+(25-25)^2\times4+(24-25)^2\times4]/10=[8+0+4]/10=1{,}2$

INDICI DI VARIABILITA' RELATIVI

Gli indici assoluti sono espressi in una unità di misura (solitamente la stessa in cui è espresso il carattere)

Ciò facilita l'interpretazione del valore numerico ma impedisce il confronto tra la variabilità di caratteri espressi in unità di misure differenti

Inoltre può essere improprio confrontare con un indice assoluto la variabilità di due distribuzioni riferite a caratteri che, seppure espressi nella stessa unità di misura, hanno livelli molto diversi. Es: variabilità del peso delle madri e dei neonati (entrambi in kg) o la variabilità dei redditi dei residenti negli USA e nello Yemen (entrambi in $).

Relativizzazione rispetto ad un valor medio

Coefficiente di Variazione $CV = \frac{\sigma}{\overline{x}}$

Non dipende dall'unità di misura
Non dipende dall'ordine di grandezza della variabile

CVmin=0 **CVmax=??**

Relativizzazione rispetto ad un valor massimo Rapporto di Concentrazione

Problema: come definire il massimo?

Soluzione: Nel caso di caratteri additivi, non negativi, trasferibili si può definire una distribuzione massimizzante la variabilità

Si consideri la successione di modalità osservate del carattere X ordinate in senso non decrescente

$$(x_1 \le x_2 \le \ldots \le x_j, \le \ldots \le x_{N-1} \le x_N)$$

Sia M la media aritmetica di X. Il totale ottenuto sommando le modalità osservate (ammontare complessivo del carattere) è NM

	distribuzioni nelle quali la varia-bilità è quella		
unità	**min**	**osservata**	**max**
1	M	x_1	0
2	M	x_2	0
3	M	x_3	0
…	…	…	…
j	M	…	0
…	…	…	…
N-2	M	x_{N-2}	0
N-1	M	x_{N-1}	0
N	M	x_N	NM
totale	***NM***	***NM***	***NM***

L’ammontare complessivo del carattere è sempre lo stesso nelle tre situazioni (successioni), quella osservata, quella min e quella max. Cambia la distribuzione interna: nel caso min si parla di equidistribuzione; nel caso max di massima concentrazione

Concentrazione → aspetto particolare della variabilità

L’indice che si utilizza per misurare la concentra zione è il rapporto di concentrazione (R) definito come

$$R = \frac{\Delta}{2M} = \frac{\sum_i \sum_j |x_i - x_j|}{N(N-1)} \cdot \frac{1}{2M}$$

Dove 2M è il valore che raggiunge la differenza media semplice (Δ) nel caso di massima concentrazione

Infatti, in caso di massima concentrazione la matrice di tutte le possibili differenze diviene

	$x_1=0$	$x_2=0$	….	$x_j=0$	….	$x_{N-1}=0$	$x_N=NM$
$x_1=0$	0	0	0	0	0	0	|0-NM|
$x_2=0$	0	0	0	0	0	0	|0-NM|
….	0	0	0	0	0	0	|0-NM|
$x_j=0$	0	0	0	0	0	0	|0-NM|
….	0	0	0	0	0	0	|0-NM|
$x_{N-1}=0$	0	0	0	0	0	0	|0-NM|
$x_N=NM$	|NM-0|	|NM-0|	|NM-0|	|NM-0|	|NM-0|	|NM-0|	0

e la differenza media è

$$\Delta_{max} = \frac{\sum_i \sum_j |x_i - x_j|}{N(N-1)} = \frac{2MN(N-1)}{N(N-1)} = 2M$$

Es: Numero colli di un prodotto giacenti nei 10 magazzini di una ditta nel 2014

mag	2014
m1	230
m2	150
m3	10
m4	20
m5	100
m6	200
m7	50
m8	10
m9	20
m10	210
totale	1000

Giacenza media
M=1000/10=100

	10	10	20	20	50	100	150	200	210	230
10	0	0	10	10	40	90	140	190	200	220
10	0	0	10	10	40	90	140	190	200	220
20	10	10	0	0	30	80	130	180	190	210
20	10	10	0	0	30	80	130	180	190	210
50	40	40	30	30	0	50	100	150	160	180
100	90	90	80	80	50	0	50	100	110	130
150	140	140	130	130	100	50	0	50	60	80
200	190	190	180	180	150	100	50	0	10	30
210	200	200	190	190	160	110	60	10	0	20
230	220	220	210	210	180	130	80	30	20	0

9440

Δ = 9440/(10*9)=104,889

R= Δ/2M = 104,889/200 = 0,524

DIAGRAMMA DI CONCENTRAZIONE *(DI LORENZ)*

SiA X il carattere reddito rilevato su N individui e si consideri la successione ordinata in senso non decrescente

$$(x_1 \le x_2 \le \dots \le x_j, \le \dots \le x_{N-1} \le x_N)$$

Sia M la media aritmetica di X. Il totale ottenuto sommando le modalità osservate (ammontare complessivo del carattere) è NM

$p_j = \frac{j}{N}$ ***frazione di redditieri che hanno un reddito ≤ di x_j***

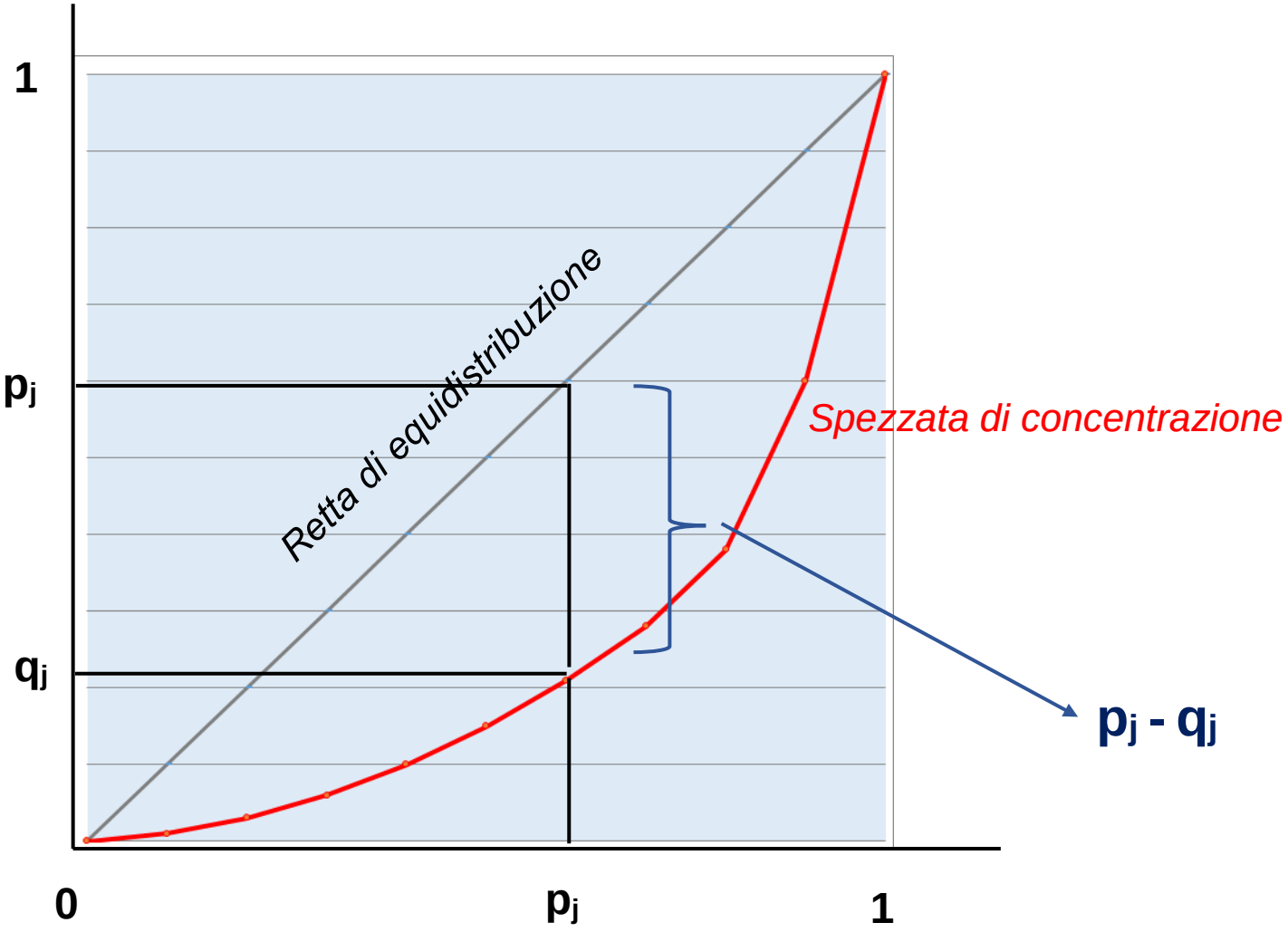
$q_j = \frac{x_1 + x_2 + \cdots + x_j}{NM}$ ***frazione del reddito totale posseduto da quei redditieri***

equidistribuz→ $q_j = \frac{JM}{NM} = \frac{JM}{NM} = p_j$

max concentraz→ $q_1=0, q_2=0, \dots, q_{N-1}=0, q_N=1$

situazione intermedia→ $q_j < p_j \quad \forall j=1,2, \dots, N-1$

Diagramma di concentrazione rappresentaz. su un piano cartesiano delle coppie (p_j , q_j)



Una misura di concentrazione

$$R^* = \frac{\sum_{j=1}^{N-1}(p_j - q_j)}{\sum_{J=1}^{N-1} p_j}$$

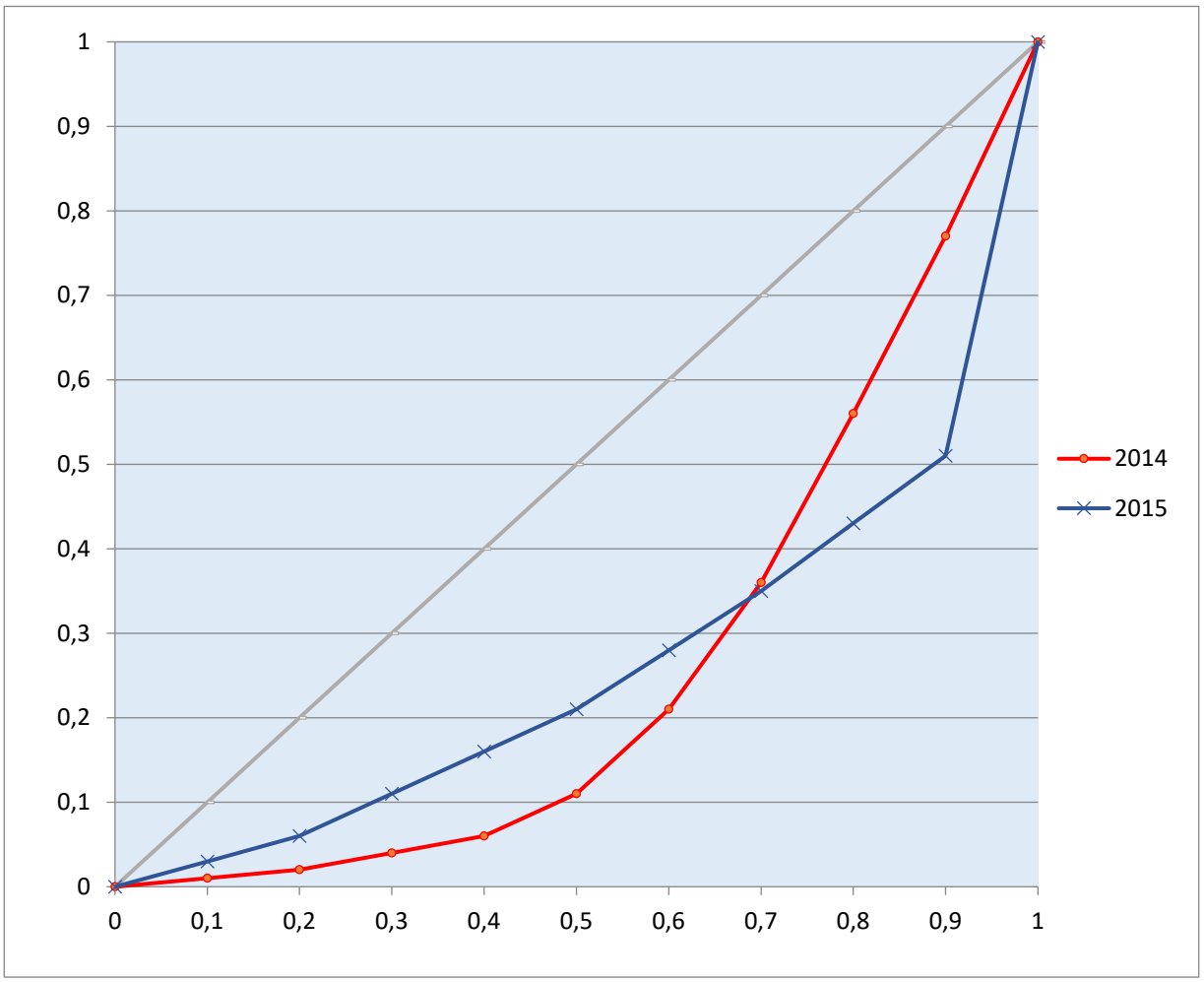
Si dimostra che

$$R^* = \frac{\Delta}{2M} = R$$

Merce giacente in magazzino nel 2014 e nel 2015

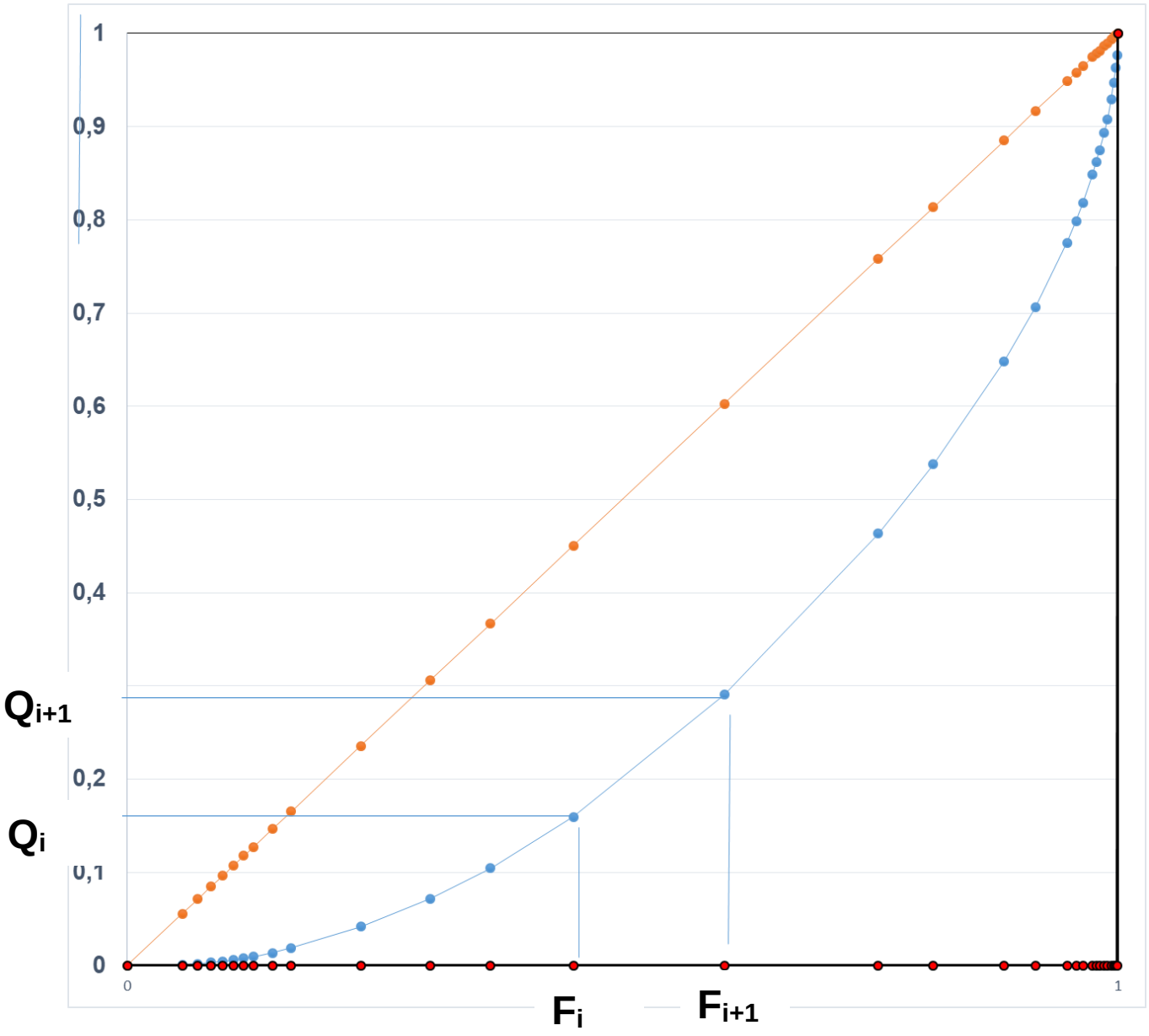
mag	2014	2015
m1	230	735
m2	150	120
m3	10	45
m4	20	120
m5	100	75
m6	200	105
m7	50	75
m8	10	75
m9	20	45
m10	210	105
tot	1000	1500

				2014	*2015*	*2014*	*2015*
	dati ordinati		Pi	Qi	Qi	Pi - Qi	Pi - Qi
m1	10	45	0,1	0,01	0,03	0,09	0,07
m2	10	45	0,2	0,02	0,06	0,18	0,14
m3	20	75	0,3	0,04	0,11	0,26	0,19
m4	20	75	0,4	0,06	0,16	0,34	0,24
m5	50	75	0,5	0,11	0,21	0,39	0,29
m6	100	105	0,6	0,21	0,28	0,39	0,32
m7	150	105	0,7	0,36	0,35	0,34	0,35
m8	200	120	0,8	0,56	0,43	0,24	0,37
m9	210	120	0,9	0,77	0,51	0,13	0,39
m10	230	735	1	1	1		
tot	1000	1500	4,5			2,36	2,36
					R=	0,524	0,524



caratteri con modalità raggruppate in classi?

Classi reddito	Frequenza	Ammontare (migliaia)	freq cumul	ammontare cumul	Fi	Qi	Fi - Fi-1	Qi+Qi+1	(Fi - Fi-1) * (Qi+Qi+1)
da 0 a 1000	2.234.286	956.394	2.234.286	956.394	0,056	0,001	0,056	0,001	6,52E-05
da 1000 a 1500	612.343	760.480	2.846.629	1.716.874	0,071	0,002	0,015	0,003	4,99E-05
da 1500 a 2000	535.423	933.470	3.382.052	2.650.344	0,085	0,003	0,013	0,005	7,13E-05
da 2000 a 2500	478.121	1.077.166	3.860.173	3.727.510	0,096	0,005	0,012	0,008	9,30E-05
da 2500 a 3000	448.283	1.231.765	4.308.456	4.959.275	0,108	0,006	0,011	0,011	1,19E-04
da 3000 a 3500	401.256	1.303.330	4.709.712	6.262.605	0,118	0,008	0,010	0,014	1,37E-04
da 3500 a 4000	391.474	1.469.114	5.101.186	7.731.719	0,127	0,009	0,010	0,017	1,67E-04
da 4000 a 5000	769.734	3.472.581	5.870.920	11.204.300	0,147	0,014	0,019	0,023	4,45E-04
da 5000 a 6000	749.011	4.126.665	6.619.931	15.330.965	0,165	0,019	0,019	0,032	0,000606
da 6000 a 7500	2.816.050	18.894.841	9.435.981	34.225.806	0,236	0,042	0,070	0,060	0,004256
da 7500 a 10000	2.821.276	24.661.851	12.257.257	58.887.657	0,306	0,072	0,070	0,114	0,008012
da 10000 a 12000	2.421.347	26.590.132	14.678.604	85.477.789	0,367	0,104	0,061	0,176	0,010661
da 12000 a 15000	3.341.557	45.180.558	18.020.161	130.658.347	0,450	0,159	0,083	0,264	0,022027
da 15000 a 20000	6.104.263	107.445.327	24.124.424	238.103.674	0,603	0,291	0,153	0,450	0,068653
da 20000 a 26000	6.224.701	141.641.787	30.349.125	379.745.461	0,758	0,464	0,156	0,754	0,117296
da 26000 a 29000	2.215.340	60.780.191	32.564.465	440.525.652	0,814	0,538	0,055	1,001	0,055422
da 29000 a 35000	2.864.038	90.707.170	35.428.503	531.232.822	0,885	0,648	0,072	1,186	0,084883
da 35000 a 40000	1.265.383	47.148.534	36.693.886	578.381.356	0,917	0,706	0,032	1,354	0,042823
da 40000 a 50000	1.280.775	56.804.654	37.974.661	635.186.010	0,949	0,775	0,032	1,481	0,047405
da 50000 a 55000	363.838	19.056.998	38.338.499	654.243.008	0,958	0,799	0,009	1,574	0,014308
da 55000 a 60000	275.036	15.785.413	38.613.535	670.028.421	0,965	0,818	0,007	1,616	0,011108
da 60000 a 70000	393.471	25.446.995	39.007.006	695.475.416	0,975	0,849	0,010	1,667	0,016387
da 70000 a 75000	150.781	10.922.011	39.157.787	706.397.427	0,978	0,862	0,004	1,711	0,006447
da 75000 a 80000	126.131	9.764.434	39.283.918	716.161.861	0,982	0,874	0,003	1,736	0,005472
da 80000 a 90000	186.460	15.782.982	39.470.378	731.944.843	0,986	0,893	0,005	1,768	0,008235
da 90000 a 100000	127.398	12.067.189	39.597.776	744.012.032	0,989	0,908	0,003	1,802	0,005735
da 100000 a 120000	155.153	16.897.454	39.752.929	760.909.486	0,993	0,929	0,004	1,837	0,007121
da 120000 a 150000	113.381	15.092.973	39.866.310	776.002.459	0,996	0,947	0,003	1,876	0,005315
da 150000 a 200000	77.244	13.195.052	39.943.554	789.197.511	0,998	0,963	0,002	1,911	0,003687
da 200000 a 300000	46.696	11.161.576	39.990.250	800.359.087	0,999	0,977	0,001	1,940	0,002264
oltre 300000	31.772	18.893.998	40.022.022	819.253.085	1,000	1,000			
TOTALE	40.022.022	819.253.085							0,549271



Rapporto di concentrazione come rapporto tra area di concentrazione

$$Area\ conc = \frac{1}{2} - \frac{\sum_{i=1}^{k-1}(F_{i+1} - F_i)(Q_{i+1} + Q_i)}{2}$$

e area di max concentraz = ½

$$R = 1 - \sum_{i=1}^{k-1}(F_{i+1} - F_i)(Q_{i+1} + Q_i)$$

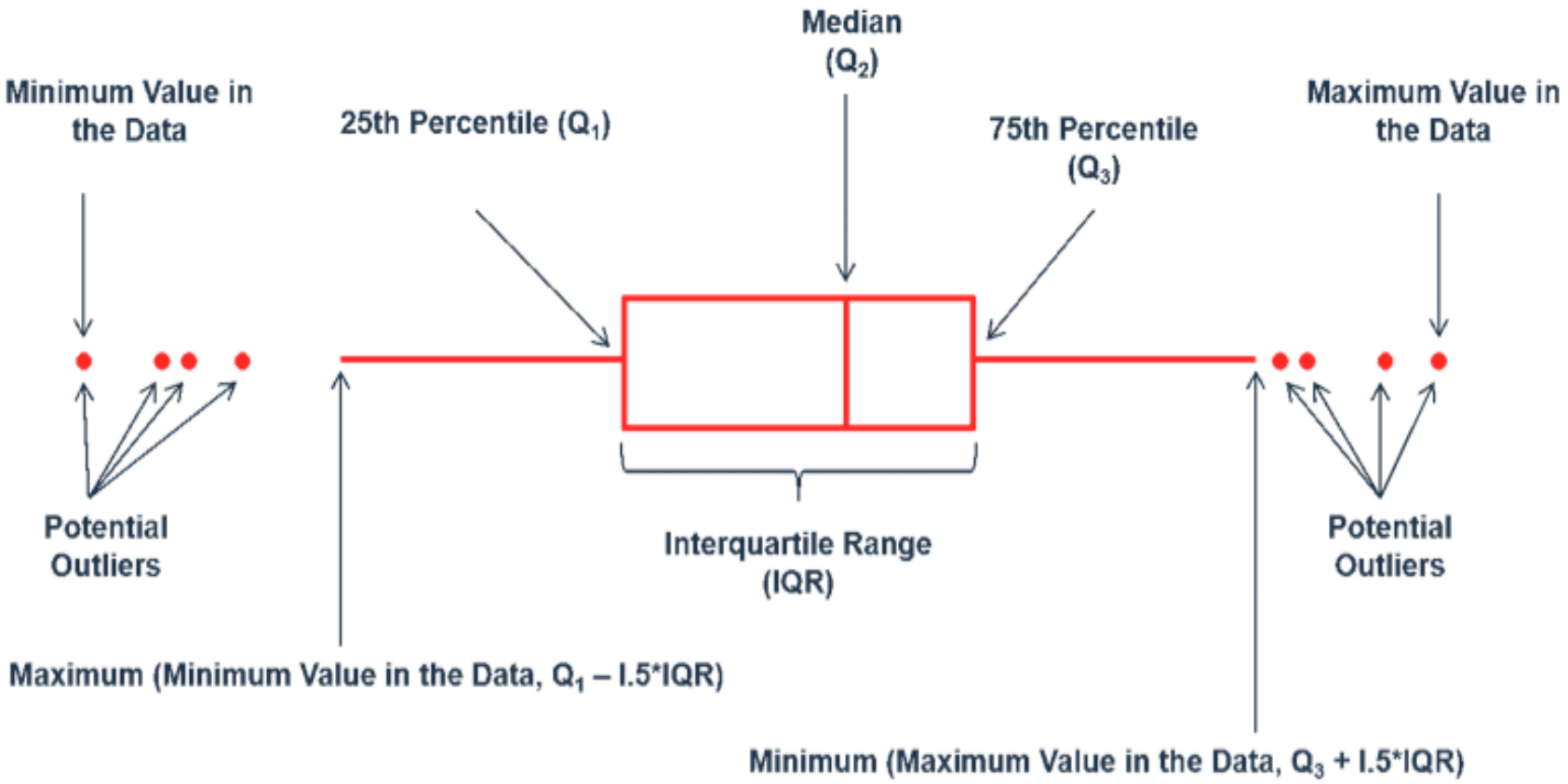
1 - 0,549271 = 0.450729

una domanda

classi di reddito	paese A n° reddi-tieri	paese B n° reddi-tieri
100-200	25	5
200-400	25	5
400-600	25	85
600-800	25	5

il reddito presenta maggiore concentrazione nel paese A o nel paese B?

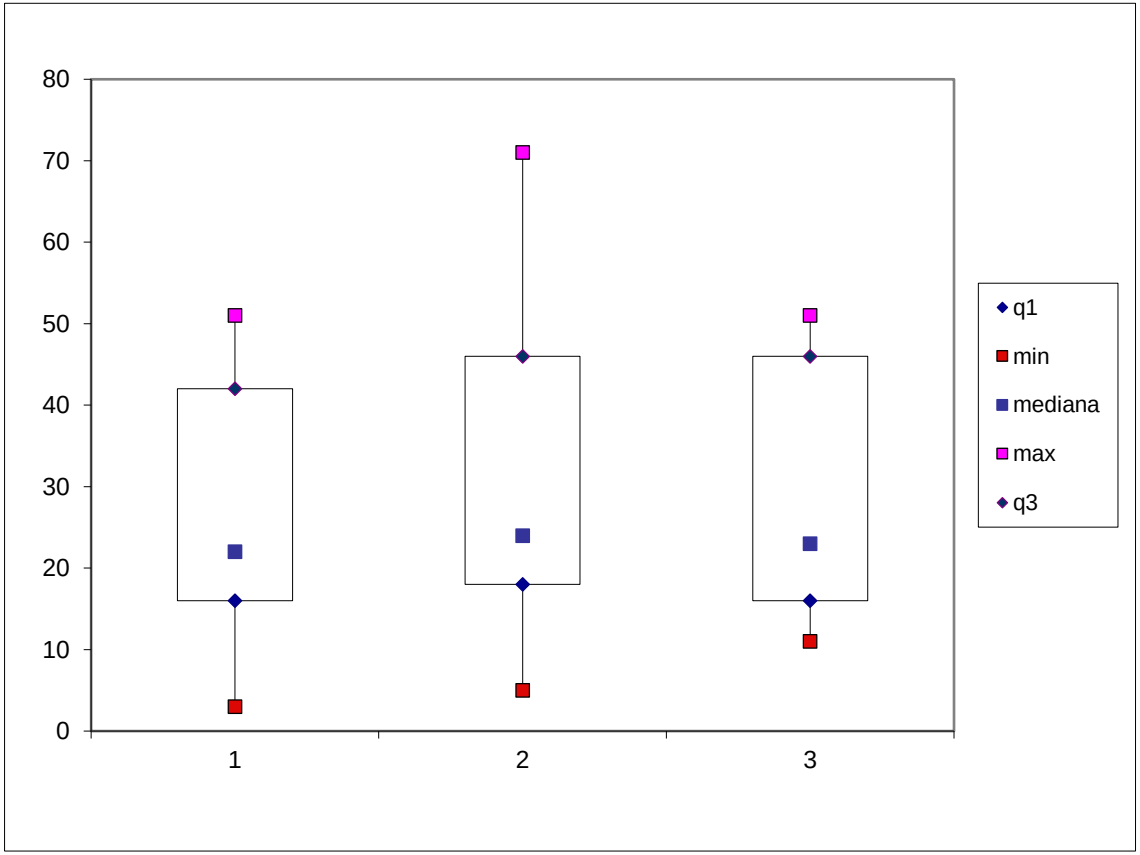
BOX PLOT - UN GRAFICO CHE SINTETIZZA LE INFORMAZIONI SU VALORI MEDI E VARIABILITA'

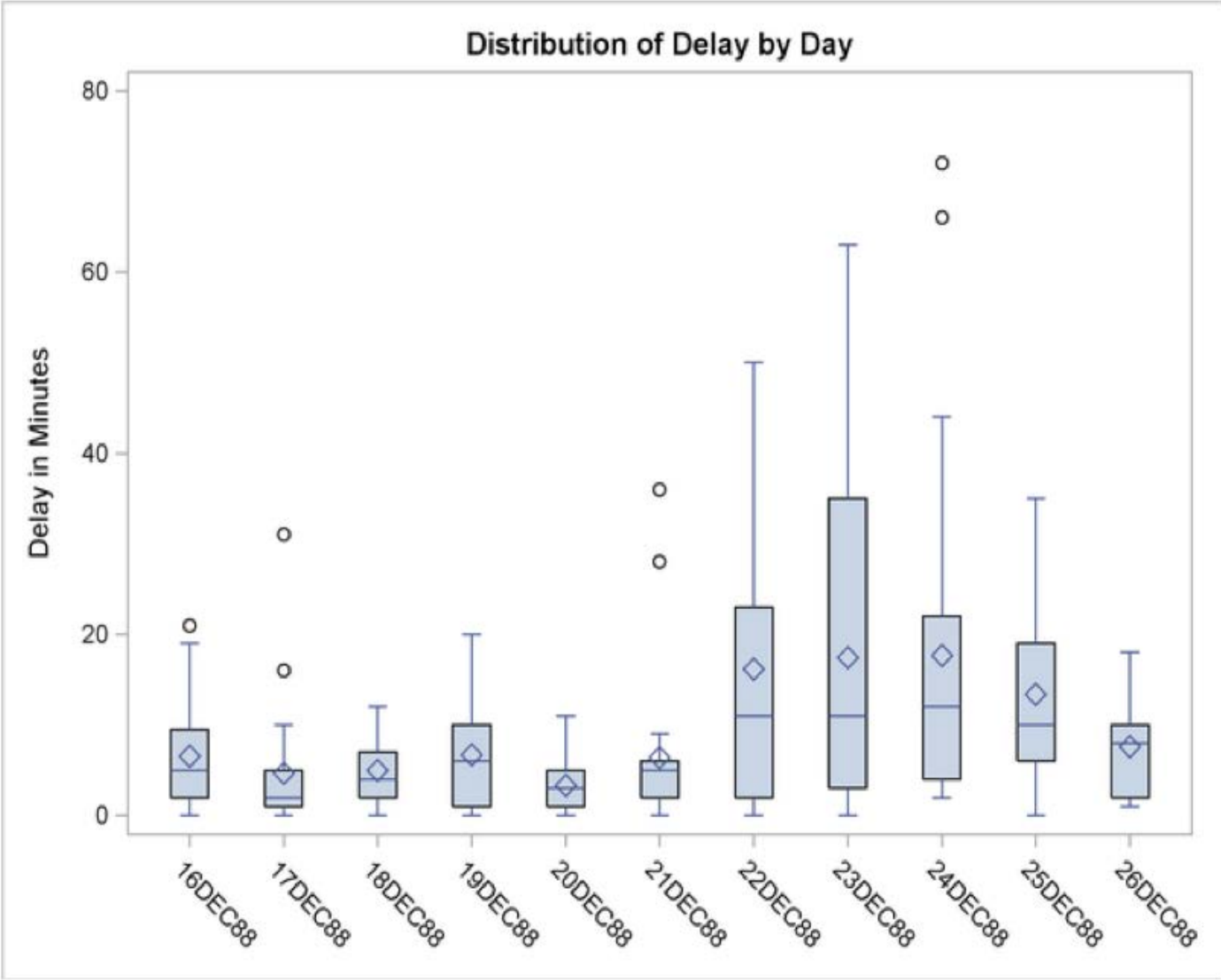


	Età anni compiuti		
	C1	C2	C3
1	16	18	16
2	41	18	22
3	43	20	47
4	7	43	51
5	16	42	17
6	22	71	22
7	49	17	24
8	49	42	47
9	16	47	51
10	40	17	16
11	38	49	42
12	16	46	44
13	3	24	11
14	17	17	16
15	51	22	45
16	43	25	39
17	16	56	11
18	42	50	11
19	39	20	17
20	8	22	50
21	4	8	49
22	16	5	23
23	4	39	13
24	45	17	16
25	39	47	46

dati ordinati

	C1	C2	C3
min	3	5	11
q1	16	18	16
mediana	22	24	23
q3	42	46	46
max	51	71	51




Distribution of Delay by Day
Delay in Minutes
80
60
40
20
0
16DEC88
17DEC88
18DEC88
19DEC88
20DEC88
21DEC88
22DEC88
23DEC88
24DEC88
25DEC88
26DEC88