

Statistica

Maura Mezzetti

maura.mezzetti@uniroma2.it

Statistica Descrittiva

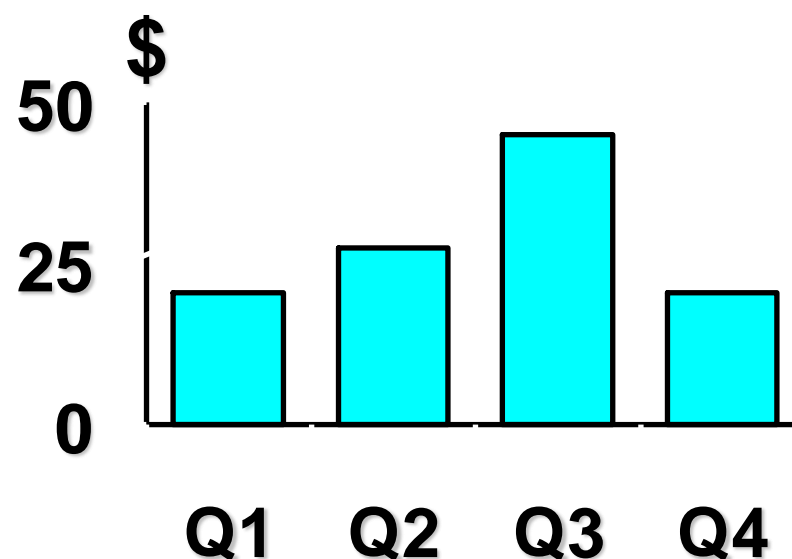
1° parte

Statistica Descrittiva

- Caratteri e scale di misura
- La distribuzione di un carattere
- La distribuzione di un carattere: le medie e la variabilità
- Analisi dell'associazione tra due caratteri

Statistica Descrittiva

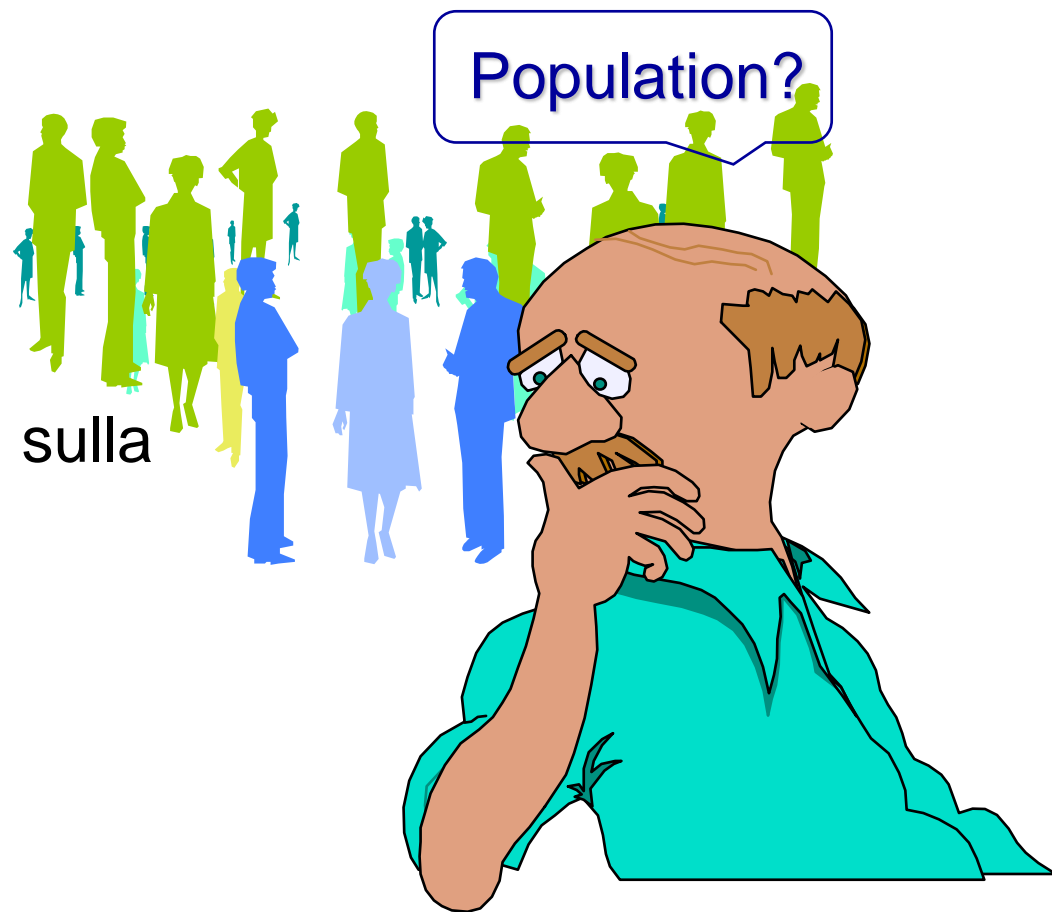
- Consiste in:
 - Raccolta dati
 - Presentazione dei dati
 - Sintesi dei dati
- Scopo
 - Descrizione dei dati



$$\bar{X} = 30.5 \quad S^2 = 113$$

Statistica Inferenziale

- Consiste in:
 - Stima
 - Verifica di ipotesi
- Scopo:
 - Prendere decisioni sulla popolazione



Le Fonti dei Dati

- **Rilevazione di dati**: complesso di operazioni con le quali si perviene alla conoscenza dei dati, ossia delle modalità di uno o più caratteri di un collettivo statistico.

Fonti statistiche ufficiali

- Il Sistema Statistico Nazionale, SISTAN, viene istituito nel 1989 e altro non è che una rete di circa diecimila operatori pubblici e privati, operanti in circa 3500 uffici, preposti, in base alle norme vigenti, a fornire le statistiche ufficiali nazionali. Il suo obiettivo è coordinare tutte le competenze e le attività di raccolta dei dati nei vari organismi centrali e periferici della pubblica amministrazione.

ISTAT

Il coordinamento del Sistan è affidato all'Istituto nazionale di statistica (Istat), posto sotto la vigilanza del Presidente del Consiglio, a cui spettano i seguenti compiti:

- Coordinare il Sistema Statistico Nazionale;
- Predisporre il Programma Statistico Nazionale (PSN);
- Predisporre nomenclature e metodologie ufficiali e vincolanti;
- Diffondere i dati delle indagini effettuate;
- Mantenere rapporti con enti statistici esteri ed internazionali

Il sistema statistico europeo

- Nell'Unione Europea, le statistiche ufficiali sono affidate all'**EUROSTAT**, il cui compito è raccogliere ed elaborare i dati statistici riguardanti i paesi comunitari e i principali partners commerciali. L'ufficio statistico dell'Unione Europea svolge, più che funzioni di produzione, compiti di coordinamento e di definizione di standard comuni fra gli uffici statistici dei paesi dell'Unione

Alcune fonti internazionali

- L'Organizzazione per la cooperazione e lo sviluppo economico (OCSE).
- L'Organizzazione delle Nazioni Unite (ONU) e gli enti ad esso collegati.
- L'Organizzazione per l'alimentazione e l'agricoltura (FAO).
- L'Ufficio Internazionale del Lavoro (BIT) per le statistiche del lavoro.

Alcune fonti internazionali

- L'Organizzazione Mondiale della Sanità (OMS) per le statistiche sanitarie.
- L'Organizzazione delle Nazioni Unite per l'educazione, la scienza e la cultura (UNESCO) per le statistiche dell'istruzione. Indagini internazionali periodiche su orientamenti valorali, atteggiamenti, opinione pubblica, ecc. comprendono l'Eurobarometro e le ricerche condotte nell'ambito dell'International Social Survey Program, le European Values Surveys e le World Values Surveys.

Le Fonti Statistiche Ufficiali

- **Rilevazione diretta** Rilevazioni dove l'informazione viene espressamente raccolta al fine di conoscere un determinato fenomeno
- **Esempi:** i censimenti, le indagini campionarie ad hoc su settori specifici (indagini multiscopo, forze lavoro, consumi, sugli sbocchi professionali dei laureati etc.)



Le Fonti Statistiche Ufficiali

- **Rilevazione indiretta** (o indagine statistica basata su dati di fonte amministrativa) un'indagine che utilizza i dati amministrativi con finalità di tipo statistico La gran parte delle rilevazioni svolte dall'Istat ha base amministrativa
- Esempi
 - statistiche sul commercio con l'estero a partire dalle bolle doganali,
 - statistiche dell'istruzione attraverso la registrazione degli iscritti e dei licenziati negli istituti scolastici
 - statistiche sanitarie a seguito delle registrazioni fatte dai medici e dalle strutture ospedaliere etc.

Fonti di dati

Fonti di dati nazionali

- www.istat.it
 - Microdata
- <https://www.bancaditalia.it/>
 - Statistiche

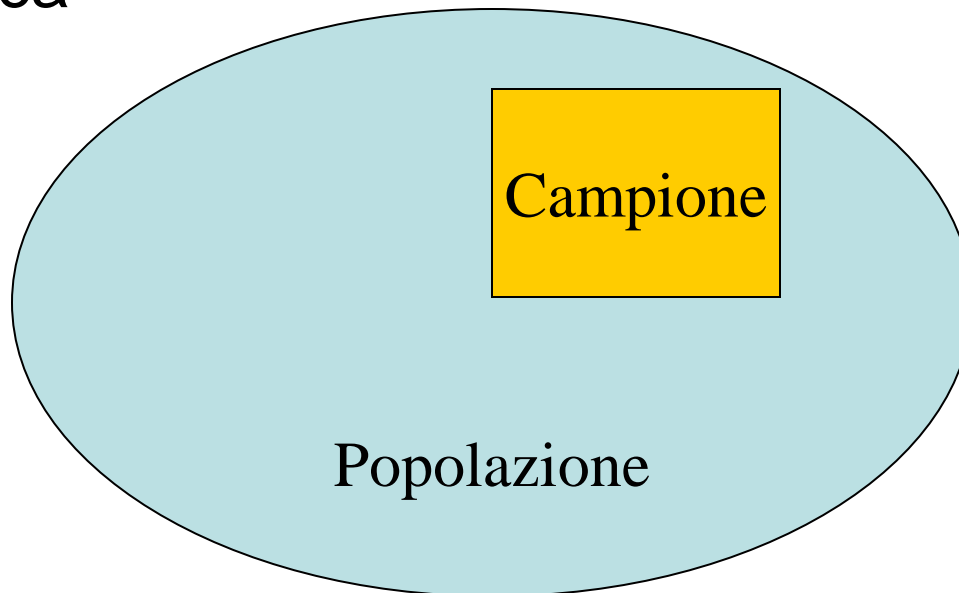
Fonti di dati internazionali

Soggetti afferenti alle Nazioni Unite e Agenzie Specializzate

- [WORLD BANK](http://www.worldbank.org)
- www.census.gov
 - <https://data.census.gov/cedsci/>

Concetti

- **Popolazione:** (o Universo) è un qualsiasi insieme di elementi che forma l'oggetto di studio di un'analisi statistica
- **Campione:** È un sotto-insieme ottenuto da una particolare popolazione e finalizzato ad un'analisi statistica

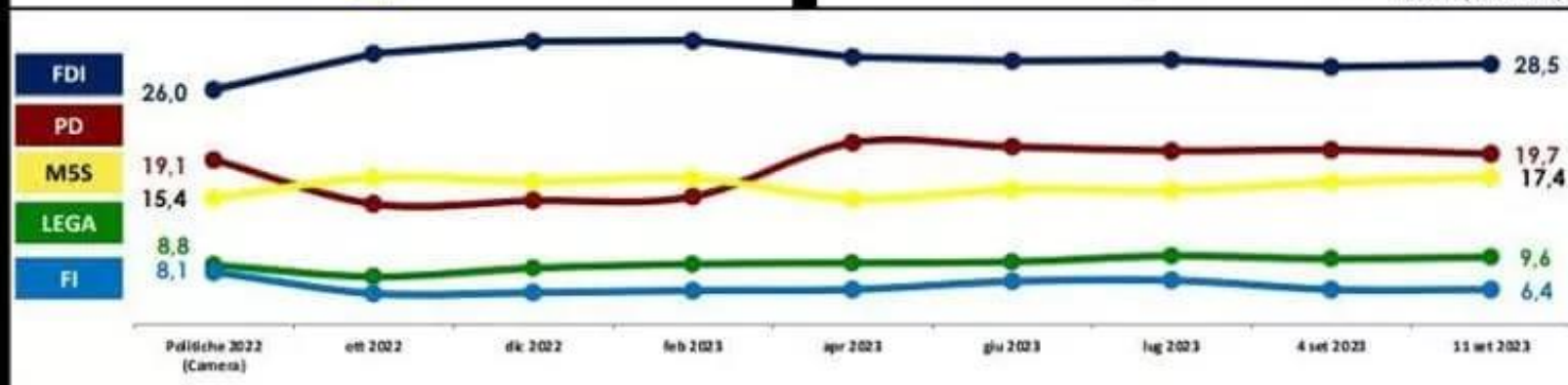


INTENZIONI DI VOTO 11 SETTEMBRE 2023

		Differenza rispetto al 4/09/2023
Fratelli d'Italia	28,5	+0,3
Partito Democratico	19,7	-0,4
Movimento 5 Stelle	17,4	+0,5
Lega	9,6	+0,2
Forza Italia	6,4	=
Azione	3,8	+0,3

		Differenza rispetto al 4/09/2023
Alleanza Verdi-Sinistra	3,3	=
+Europa	2,7	+0,1
Italia Viva	2,6	-0,2
Per l'Italia con Paragone	1,9	-0,3
Unione Popolare	1,7	-0,2
Altro partito	2,4	-0,3

Non si esprime: 39% (-1)

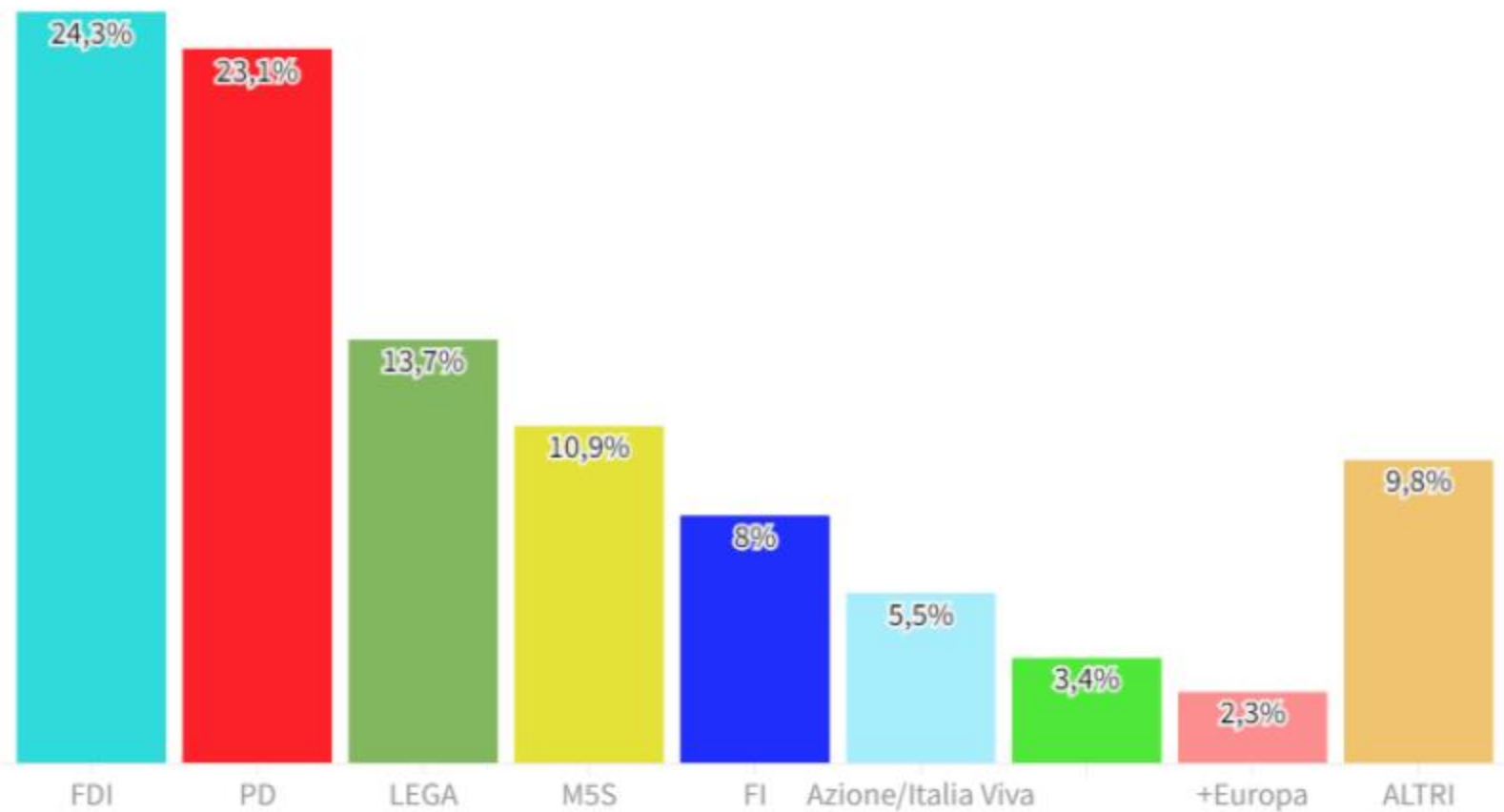


NOTA INFORMATIVA: valori espressi in %. Date di esecuzione 6 – 11 settembre 2023. Metodo di rilevazione: sondaggio CATI-CAMI-CAWI su un campione rappresentativo nazionale di 1200 soggetti maggiorenni. I dati da ottobre 2022 a luglio 2023 si riferiscono a medie mensili.

Tutti i diritti riservati 18

Media sondaggi settimanale per partiti

14 - 20 agosto



Sondaggio



- I **sondaggi elettorali** servono principalmente per avere un'idea sulle reali **intenzioni di voto** degli italiani.
- Nei **sondaggi elettorali politici** la scelta dei **parametri** sarà decisiva per avere il massimo della rappresentatività e quindi dell'affidabilità.
- Il campione deve essere indicativo su base nazionale, sarebbe quindi riduttivo effettuare l'indagine in poche regioni, così come sarebbe insufficiente condurlo solo in regioni del centro o solo in regioni del sud.

Sondaggio

Qual è la popolazione e qual è il campione?



Una **base di dati** statistica è rappresentata da un insieme di caratteristiche rilevate sulle unità statistiche che costituiscono il collettivo.

Le informazioni sono ottenute misurando su $u \in \mathcal{U}$ tali **caratteri** (caratteristiche, variabili) X .

Misurare significa attribuire una **modalità** secondo determinate regole e con certi contenuti.

$$u \mapsto x$$

x rappresenta la modalità del carattere X associata all'unità u .
Le modalità del carattere sono esaustive e mutualmente esclusive (ad ogni u si associa una e una sola modalità).

- **Unità statistica:** Rappresenta l'elemento base della popolazione, la quale può quindi essere intesa come l'insieme delle unità statistiche ad essa relative. Un'unità statistica può consistere in un individuo, un oggetto, un animale. ecc.
- **Carattere:** È il fenomeno oggetto di studio, rilevato sulle unità statistiche della popolazione di riferimento e codificato secondo le esigenze dell'analisi statistica.
- **Modalità:** È l'espressione concreta con la quale la variabile si manifesta nelle unità statistiche. La modalità può consistere in un numero (l'età di un particolare individuo) così come in una qualità (il genere di un individuo).

Distribuzioni univariate e multivariate

L'informazione statistica di base (output del processo di misurazione) prende solitamente la forma di una **distribuzione unitaria semplice** (univariata) o **multipla** (multivariata), a seconda del numero di caratteri misurati sulle unità.

Distribuzione unitaria semplice:

Unità	Modalità di X
u_1	x_1
u_2	x_2
\vdots	\vdots
u_j	x_j
\vdots	\vdots
u_n	x_n

Distribuzione unitaria multivariata:

Unità	Modalità di X	Modalità di Y	...	Modalità di Z
u_1	x_1	y_1	...	z_1
u_2	x_2	y_2	...	z_1
\vdots	\vdots	\vdots	...	\vdots
u_j	x_j	y_j	...	z_j
\vdots	\vdots	\vdots	...	\vdots
u_n	x_n	y_n	...	z_n

Esempi data set che
utilizzeremo

[illegible]

Data set:

student_survey.txt

student_survey.xls

Intervistati 60 student appena laureate in Scienze
politiche all'University of Florida

subject	gen	age	high	coll	tv	veg	party	ideology	abor
1	m	32	2.2	3.5	3	n	r	6	n
2	f	23	2.1	3.5	15	y	d	2	y
3	f	27	3.3	3.0	0	y	d	2	y
4	f	35	3.5	3.2	5	n	i	4	y
5	M	23	3.1	3.5	6	n	i	1	y

Indagine su 60 student laureati in Scienze Politiche

- *GE* = sesso (m/g)
- *AG* = anni compiuti
- *HI* = voto finale alle superiori (in scala da 1-4)
- *CO* = voto finale al college (in scala da 1 a 4)
- *DH* = distanza (in miglia) del college dalla città di residenza
- *DR* = distanza (in miglia) della classe dalla residenza attuale
- *TV* = tempo medio (in ore) passato davanti alla TV alla settimana
- *SP* = tempo medio (in ore) dedicato all'attività fisica alla settimana
- *NE* = numero di volte alla settimana leggi un giornale
- *VE* = vegetarian (yes, no),
- *PA* = affiliazione politica (D = Democrat, R = Republican, I = independent)
- *PI* = ideologia politica (1 = very liberal, 2 = liberal, 3 = slightly liberal, 4 = moderate, 5 = slightly conservative, 6 = conservative, 7 = very conservative),
- *RE* = frequenza in cui si assiste a una cerimonia religiosa (0 = never, 1 = occasionally, 2 = most weeks, 3 = every week),

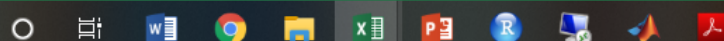
O11

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	subj	ge	ag	hi	co	dh	dr	tv	sp	ne	ve	pa	pi	re					
2	1	m	32	2,2	3,5	0	5	3	5	0	n	r	6	2					
3	2	f	23	2,1	3,5	1200	0,3	15	7	5	y	d	2	1					
4	3	f	27	3,3	3	1300	1,5	0	4	3	y	d	2	2					
5	4	f	35	3,5	3,2	1500	8	5	5	6	n	i	4	1					
6	5	m	23	3,1	3,5	1600	10	6	6	3	n	i	1	0					
7	6	m	39	3,5	3,5	350	3	4	5	7	y	d	2	1					
8	7	m	24	3,6	3,7	0	0,2	5	12	4	n	i	2	1					
9	8	f	31	3	3	5000	1,5	5	3	3	n	i	2	1					
10	9	m	34	3	3	5000	2	7	5	3	n	i	1	1					
11	10	m	28	4	3,1	900	2	1	1	2	y	i	3	0					
12	11	m	23	2,3	2,6	253	1,5	10	15	1	n	r	5	1					
13	12	f	27	3,5	3,6	190	3	14	3	7	n	d	2	1					
14	13	m	36	3,3	3,5	245	1,5	6	15	12	n	d	1	1					
15	14	m	28	3,2	3,2	500	6	3	10	1	n	i	4	1					
16	15	f	28	3	3,5	3500	1	4	3	1	n	d	1	0					
17	16	f	25	3,8	3,3	210	10	7	6	1	y	i	2	3					

Foglio1

READY

Scrivi qui per eseguire la ricerca



150%

13:53 13/09/2019

subj	ge	ag	hi	co	dh	dr	tv	sp	ne	ve	pa	pi	re
1	m	32	2,2	3,5	0	5	3	5	0	n	r	6	2
2	f	23	2,1	3,5	1200	0,3	15	7	5	y	d	2	1
3	f	27	3,3	3	1300	1,5	0	4	3	y	d	2	2
4	f	35	3,5	3,2	1500	8	5	5	6	n	i	4	1
5	m	23	3,1	3,5	1600	10	6	6	3	n	i	1	0
6	m	39	3,5	3,5	350	3	4	5	7	y	d	2	1
7	m	24	3,6	3,7	0	0,2	5	12	4	n	i	2	1
8	f	31	3	3	5000	1,5	5	3	3	n	i	2	1
9	m	34	3	3	5000	2	7	5	3	n	i	1	1
10	m	28	4	3,1	900	2	1	1	2	y	i	3	0
11	m	23	2,3	2,6	253	1,5	10	15	1	n	r	5	1
12	f	27	3,5	3,6	190	3	14	3	7	n	d	2	1
13	m	36	3,3	3,5	245	1,5	6	15	12	n	d	1	1
14	m	28	3,2	3,2	500	6	3	10	1	n	i	4	1
15	f	28	3	3,5	3500	1	4	3	1	n	d	1	0
16	f	25	3,8	3,3	210	10	7	6	1	y	i	2	3
17	f	41	4	3	1000	15	6	7	3	n	i	3	3
18	m	50	3,8	3,8	0	3	5	9	6	n	d	2	0
19	m	71	4	3,5	5000	3	6	12	2	n	i	2	0
20	f	28	3	3,8	120	1	25	0	0	y	d	1	1
21	f	26	3,7	3,7	8000	8	4	4	4	n	i	4	1
22	f	27	4	3,7	2	2,5	4	2	7	n	i	2	1
23	m	31	2,7	3,5	1700	5	7	7	2	n	r	7	3
24	f	23	3,7	3,7	2	2	7	4	2	n	i	4	0
25	m	23	3,2	3,8	450	4	0	7	7	n	i	1	0
26	f	44	3	3	0	2	2	3	2	y	i	3	2
27	m	26	3,7	3	1000	3	8	2	7	n	d	2	1

Scale di misura

- Dati Quantitativi
 - Scala discreta
 - Continua
- Dati Qualitativi
 - Scala nominale (dati categorici)
 - Scala ordinale

Variabili/caratteri qualitative e quantitative

- Le variabili statistiche possono essere **qualitative**, se esprimono una qualità dell'individuo (ad esempio colore degli occhi o dei capelli). Una variabile qualitativa non viene misurata, ma classificata in categorie sulla base delle modalità con cui essa si presenta (neri, castani, rossi, biondi).
- D'altra parte esistono le variabili/caratteri **quantitativi**, che possono essere misurate su una scala discreta (numero di carte di credito possedute, numero di dipendenti di un'azienda) o su una scala continua (reddito).

Variabili qualitative

Le modalità utilizzate per descrivere il fenomeno analizzato prendono la forma di aggettivi o di altre espressioni verbali. A loro volta i dati qualitativi possono essere

- nominali se non esiste nessun ordinamento naturale tra le modalità; esempi di dati sconnessi sono: il sesso, il tipo di servizio offerto da un albergo (mezza pensione/pensione completa ecc);
- ordinali nel caso in cui un ordinamento naturale esiste; esempi di dati qualitativi ordinali sono: il titolo di studio.

Quando le modalità sono solamente due (esempi vivo/morto) si parla di dati dicotomici o binari

Scala nominale

Le unità sono classificate in funzione dell'appartenenza ad una particolare classe o categoria.

Le modalità non assumono un ordine precostituito (sono sconnesse).

Confronto tra unità: criterio di identità

$$x_i = x_j \quad \text{ovvero} \quad x_i \neq x_j$$

Esempi: sesso (F-M), professione, settore di attività economica.

Scala ordinale

Le modalità hanno un ordine sequenziale.

Confronto tra unità: criterio di ordinamento

$$x_i < x_j \quad \text{ovvero} \quad x_i = x_j, \quad \text{ovvero} \quad x_i > x_j$$

Esempi: Titolo di studio (Lic. elementare, Lic. media, Lic. media sup., Laurea triennale, Laurea Mag.), gradimento (basso, medio, alto).

Variabili quantitative

Le modalità sono espresse da numeri. I dati quantitativi si suddividono a loro volta in dati

- discreti (**how many?**) quando le modalità sono esprimibili da numeri interi; provengono da un conteggio. esempi : il numero di clienti, il numero di pezzi prodotti;
- continui o reali (**how much?**) quando le modalità sono esprimibili da numeri reali; provengono da una misurazione. Esempi sono: il tempo d'attesa ad uno sportello, il peso di un manufatto.

Scala ad intervallo

Consente di confrontare l'intensità del fenomeno in unità diverse.
Tuttavia, non esiste una origine naturale e l'unità di misurazione è arbitraria.

Confronto tra unità: differenza

$$x_i - x_j$$

Esempio: temperatura (Celsius e Fahrenheit: $F = 32 + 1.8C$)

Celsius	Fahrenheit
-17.8	0.0
0.0	32.0
5.0	41.0
10.0	50.0
100.0	212.0

Scala di rapporto

A differenza della scala precedente, esiste un'origine naturale (zero assoluto) che denota l'assenza del carattere.

Confronto tra unità: oltre a $x_i - x_j$, ha senso calcolare i rapporti x_j/x_i .

Esempi: produzione, prezzi, fatturato, ordinativi, peso, numero componenti la famiglia, addetti.

Esercizio: tipologia di dati

- età
- Età all'ultimo compleanno (in anni)
- Il paziente è stato dal dentista nell'ultimo anno?
- Numero di volte un paziente è stato dal dentista nell'ultimo anno
- Titolo di studio
- Classe sociale
- Stato civile
- IQ
- Numero di persone nella famiglia
- Colore di autoveicoli
- Lunghezza del salto di una rana

Esercizio: tipologia di dati

- Numero di figli in famiglia
- Comune di residenza
- Distanza (in miglia) tra casa e scuola
- Periodo di studio necessario per preparare un esame
- Numero di persone in attesa in linea
- Numero di multe ricevute l'anno scorso
- Il peso di una bicicletta elettrica in produzione

Prestito	Durata	Valore casa	Saldo conto corrente	Reddito familiare	Stato civile	Titolo di studio	Num di figli
200	15	370	3324.05	31.91	SEPARATO	superiori	2
320	20	510	10.89	65.80	CONIUGATO	superiori	1
240	10	380	3903.87	43.26	CONIUGATO	media	2
360	25	560	4450.64	54.56	CONIUGATO	superiori	3
50	20	230	6688.03	12.10	SINGLE	media	0
250	20	560	591.10	48.62	CONIUGATO	laurea	2
240	10	450	7845.18	52.82	SINGLE	superiori	1
70	30	130	521.57	16.58	SINGLE	superiori	0
150	20	560	10436.73	43.06	CONIUGATO	superiori	2
100	20	450	762.43	29.45	SEPARATO	superiori	2

Indagine su 60 student laureati in Scienze Politiche

- *GE* = sesso (m/g)
- *AG* = anni compiuti
- *HI* = voto finale alle superiori (in scala da 1-4)
- *CO* = voto finale al college (in scala da 1 a 4)
- *DH* = distanza (in miglia) del college dalla città di residenza
- *DR* = distanza (in miglia) della classe dalla residenza attuale
- *TV* = tempo medio (in ore) passato davanti alla TV alla settimana
- *SP* = tempo medio (in ore) dedicato all'attività fisica alla settimana
- *NE* = numero di volte alla settimana leggi un giornale
- *VE* = vegetarian (yes, no),
- *PA* = affiliazione politica (D = Democrat, R = Republican, I = independent)
- *PI* = ideologia politica (1 = very liberal, 2 = liberal, 3 = slightly liberal, 4 = moderate, 5 = slightly conservative, 6 = conservative, 7 = very conservative),
- *RE* = frequenza in cui si assiste a una cerimonia religiosa (0 = never, 1 = occasionally, 2 = most weeks, 3 = every week),

Indagine su 60 student laureati in Scienze Politiche

- *GE* = QUALITATIVO NOMINALE
- *AG* = QUANTITATIVO DISCRETO
- *HI* = QUANTITATIVO DISCRETO
- *CO* = QUANTITATIVO DISCRETO
- *DH* = QUANTITATIVO CONTINUO
- *DR* = QUANTITATIVO CONTINUO
- *TV* = QUANTITATIVO CONTINUO
- *SP* = QUANTITATIVO CONTINUO
- *NE* = QUANTITATIVO DISCRETO
- *VE* = QUALITATIVO NOMINALE
- *PA* QUALITATIVO NOMINALE
- *PI* = QUALITATIVO ORDINALE
- *RE* = QUALITATIVO ORDINALE

Distribuzioni di frequenza

Ci proponiamo di sintetizzare una distribuzione di un carattere mediante tabelle e grafici.

Una distribuzione di frequenza è una rappresentazione tabellare che riporta le modalità del carattere ed il numero (assoluto, relativo, percentuale) delle unità che presentano il carattere con quelle modalità.

Contiamo le unità che presentano la stessa modalità. Questo ha significato per i caratteri **qualitativi** e **quantitativi discreti**.

Più in generale, nel caso dei caratteri quantitativi occorre suddividere i valori che la variabile può assumere in intervalli o **classi**.

x_i modalità della i-esima unità statistica

x_i^* i-esima modalità

Il giudizio di 20 ospiti in un albergo:

Scarso	Medio	Scarso	Buono
Scarso	Ottimo	Ottimo	Buono
Buono	Buono	Scarso	Medio
Ottimo	Medio	Medio	Medio
Buono	Ottimo	Scarso	Scarso

x_3 ????

x_3^* ????

x_i modalità della i-esima unità statistica

x_i^\star i-esima modalità

Il giudizio di 20 ospiti in un albergo:

Scarso	Medio	Scarso	Buono
Scarso	Ottimo	Ottimo	Buono
Buono	Buono	Scarso	Medio
Ottimo	Medio	Medio	Medio
Buono	Ottimo	Scarso	Scarso

x_3 Scarso

x_3^\star Buono

Distribuzione di frequenze assolute

La frequenza assoluta di una modalità rappresenta il numero di volte che questa si presenta nel collettivo.

Scarsa efficacia di sintesi in presenza di un numero elevato di modalità.

Distribuzione di frequenze (semplice)

Modalità di X	Frequenza
x_1^\star	n_1
x_2^\star	n_2
\vdots	\vdots
x_j^\star	n_j
\vdots	\vdots
x_K^\star	n_K
Totale	n

$$n = n_1 + n_2 + \cdots + n_j + \cdots + n_K = \sum_{j=1}^K n_j$$

Distribuzioni di frequenze relative e percentuali

La frequenza relativa della modalità j del carattere è il rapporto

$$f_j = \frac{n_j}{n}, \quad j = 1, \dots, K.$$

La frequenza percentuale è:

$$p_j = 100 \cdot \frac{n_j}{n} = 100 \cdot f_j, \quad j = 1, \dots, K.$$

Ovviamente,

$$0 \leq f_j \leq 1, \quad \sum_{j=1}^K f_j = 1; \quad 0 \leq p_j \leq 100, \quad \sum_{j=1}^K p_j = 100.$$

Distribuzione di frequenze relative:

Modalità di X	Frequenza
x_1^\star	f_1
x_2^\star	f_2
\vdots	\vdots
x_j^\star	f_j
\vdots	\vdots
x_K^\star	f_K
Totale	1

Distribuzione di frequenze percentuali:

Modalità di X	Frequenza
x_1^\star	p_1
x_2^\star	p_2
\vdots	\vdots
x_j^\star	p_j
\vdots	\vdots
x_K^\star	p_K
Totale	100

Frequenze cumulate

La frequenza cumulata associata ad una modalità del carattere misura il numero dei casi che presentano un valore non superiore a quella modalità.

NB. Ha significato solo se il carattere è misurato su scala almeno ordinale.

Frequenza assoluta cumulata:

$$N_j = \sum_{k=1}^j n_k, j = 1, \dots, K$$

(n.b. $N_K = n$)

Modalità di X	Freq. ass.	Freq. cum.
x_1^\star	n_1	$N_1 = n_1$
x_2^\star	n_2	$N_2 = n_1 + n_2$
\vdots	\vdots	\vdots
x_j^\star	n_j	$N_j = n_1 + n_2 + \cdots + n_j$
\vdots	\vdots	\vdots
x_K^\star	n_K	n
Totale	n	

N.B.: vale la formula ricorsiva $N_j = N_{j-1} + n_j$

Frequenza relativa cumulata:

$$F_j = \sum_{k=1}^j f_k, j = 1, \dots, K$$

(n.b. $F_K = 1$)

Frequenza percentuale cumulata:

$$P_j = \sum_{k=1}^j p_k, j = 1, \dots, K$$

Modalità di X	Freq. rel.	Freq. rel. cum.
x_1^\star	f_1	$F_1 = f_1$
x_2^\star	f_2	$F_2 = f_1 + f_2$
\vdots	\vdots	\vdots
x_j^\star	f_j	$F_j = f_1 + f_2 + \cdots + f_j$
\vdots	\vdots	\vdots
x_K^\star	f_K	1
Totale	1	

N.B.: vale la formula ricorsiva $F_j = F_{j-1} + f_j$

Esempio

Distribuzione percentuale delle famiglie per classi di reddito familiare annuale a seconda del titolo di studio (Anno 2004, Banca d'Italia)

<i>Classi di reddito</i> (migl. euro)	<i>Famiglie</i>				
	Lic.	Elem.	Media Inf.	Media Sup.	Laurea
0 ÷ 5		1.6	1.4	0.6	0.8
5 ÷ 10		11.9	5.5	2.0	0.3
10 ÷ 15		24.4	10.3	5.2	1.8
15 ÷ 20		22.8	19.6	11.7	4.2
20 ÷ 30		25.4	27.7	26.7	16.0
30 ÷ 40		7.9	18.3	20.1	17.9
40 ÷		6.0	17.2	33.7	59.0

Rappresentazione grafica per variabili qualitative

- Le due rappresentazioni grafiche principali per sintetizzare una variabile qualitative sono:
 - **Diagramma a torta**: un cerchio in cui a ciascuna modalità corrisponde uno “spicchio di torta”. L’ampiezza di ogni fetta corrisponde alla percentuale che compete a ciascuna modalità
 - **Grafico a Barre**: mostra delle barre verticali di uguale base per ogni categoria. L’altezza di ciascun rettangolo è la percentuale di ogni modalità. I rettangoli sono di solito uniformemente distanziati.

Pie Charts

- Pie charts:
 - used for summarizing a categorical variable
 - Drawn as a circle where each category is represented as a “slice of the pie”
 - The size of each pie slice is proportional to the percentage of observations falling in that category

Titolo di studio dei genitori

Classe sociale di provenienza

■ entrambi i genitori con laurea
 ■ un solo genitore con la laurea
 ■ nessun genitore laureato

Agrario-veterinario

Architettura e ing. civile

Arte e design

Economico

Educazione e formazione

Giuridico

Informatica e tec.ICT

Ing. industriale e dell'info



Lett.-umanistico

Linguistico

Medico e farmaceutico

Politico-sociale e comun.

Psicologico

Scientifico

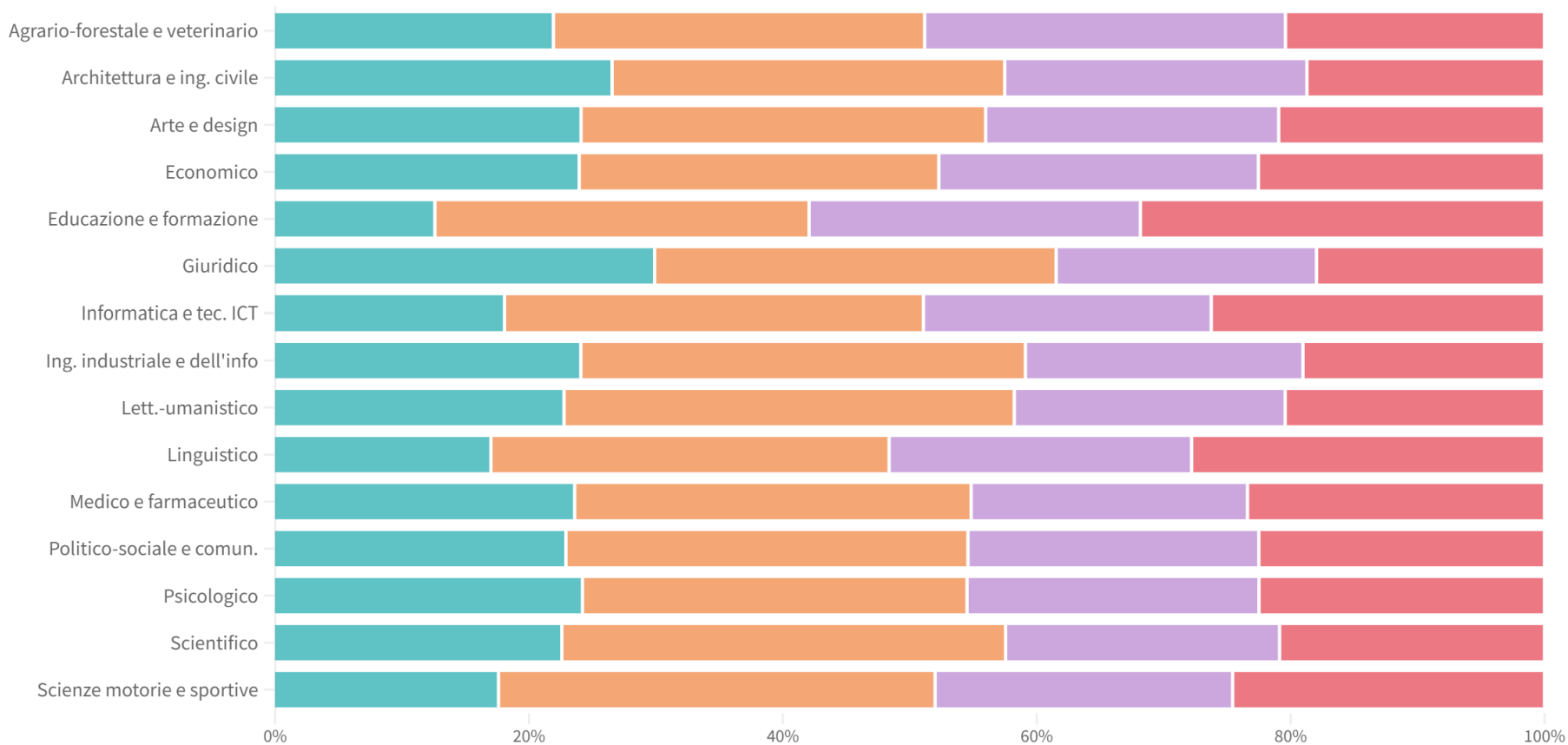
Scienze motorie e sport



Titolo di studio dei genitori

Classe sociale di provenienza

■ elevata ■ media impiegatizia ■ media autonoma ■ da lavoro esecutivo

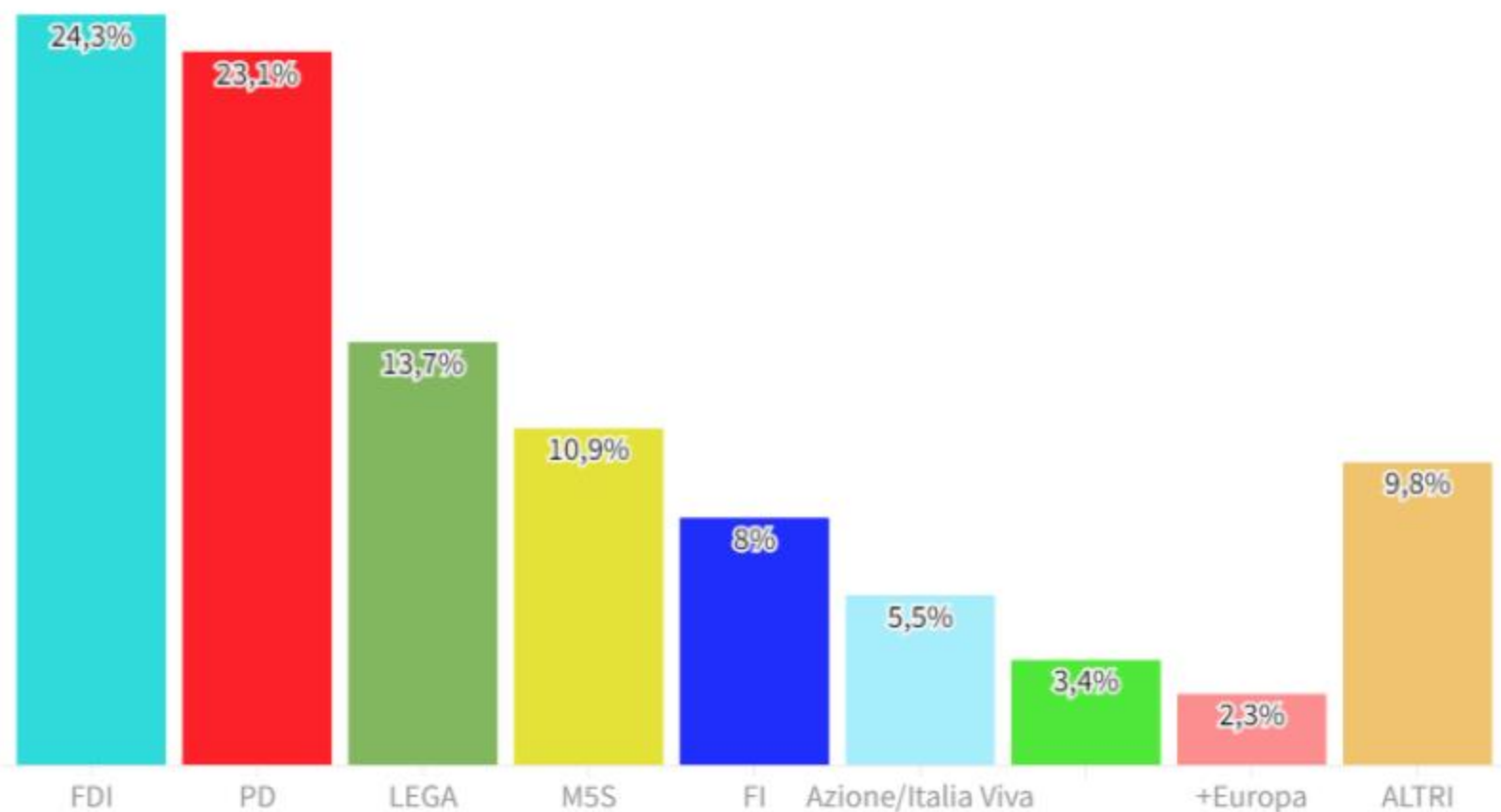


Bar Graphs

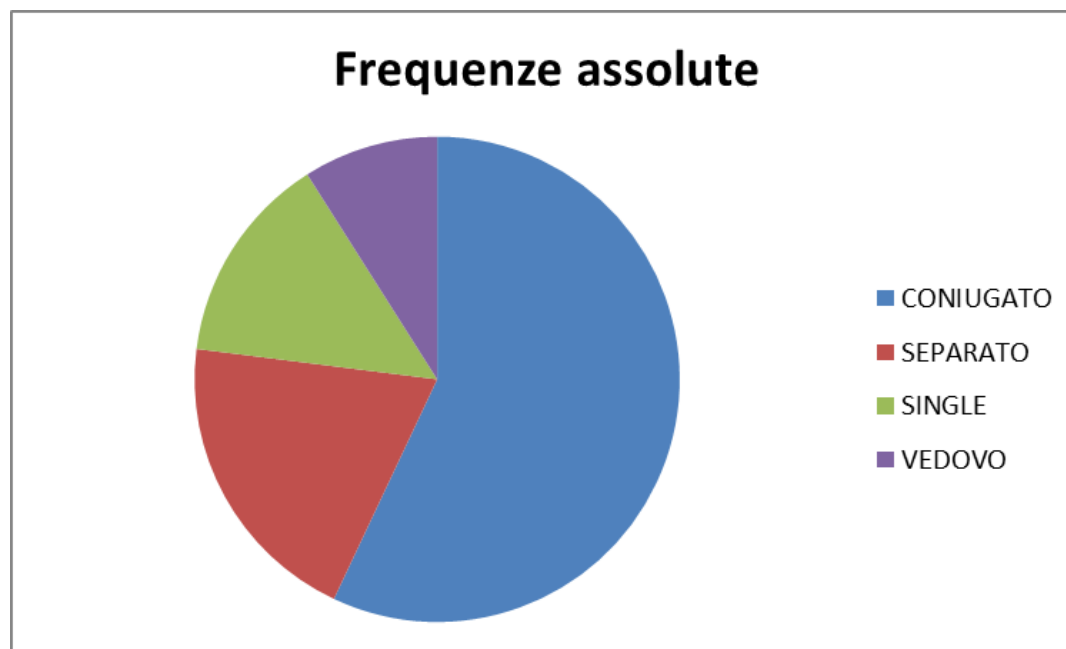
- Bar graphs are used for summarizing a categorical variable
- Bar Graphs display a vertical bar for each category
- The height of each bar represents either counts (“frequencies”) or percentages (“relative frequencies”) for that category
- Usually easier to compare categories with a bar graph than with a pie chart

Media sondaggi settimanale per partiti

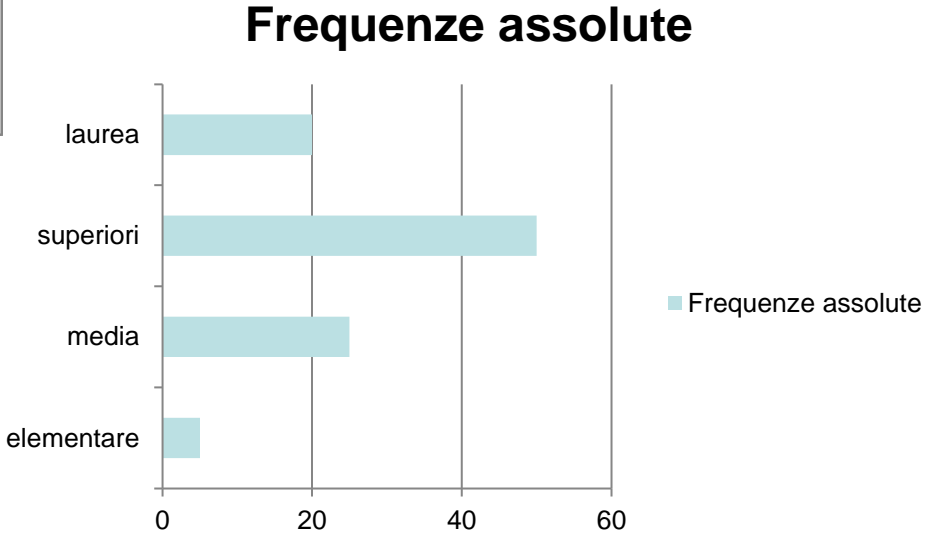
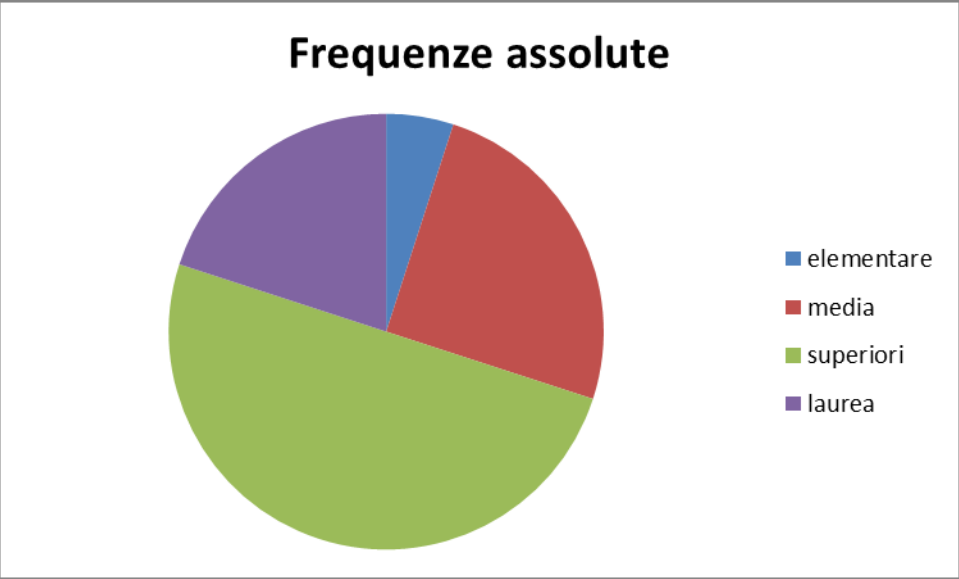
14 - 20 agosto



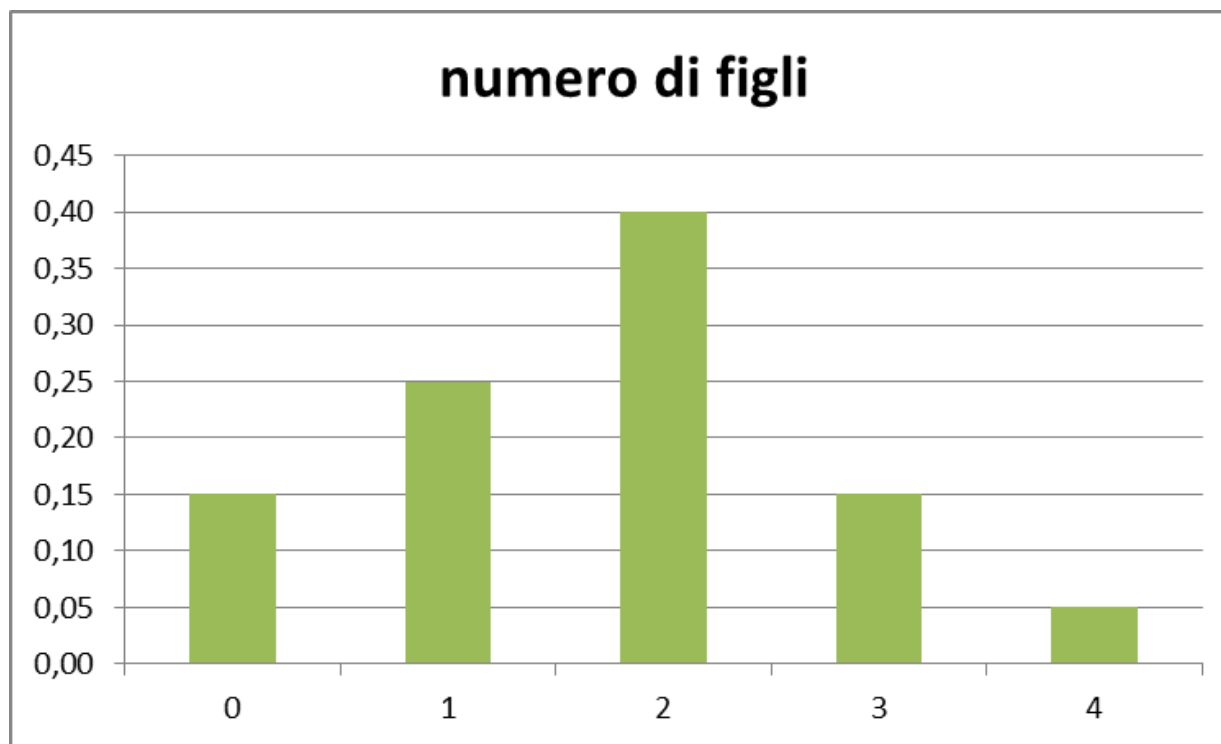
Stato civile	Frequenze assolute	Frequenze relative
CONIUGATO	57	0.57
SEPARATO	20	0.20
SINGLE	14	0.14
VEDOVO	9	0.09
Totale complessivo	100	1



Titolo di studio	Frequenze assolute	Frequenze relative	Ffrequenze cumulate
elementare	5	0.05	0.05
media	25	0.25	0.3
superiori	50	0.50	0.8
laurea	20	0.20	1
Totale complessivo	100	1	



Num di figli	Frequenze assolute	Frequenze relative	frequenze cumulate
0	15	0.15	0.15
1	25	0.25	0.4
2	40	0.4	0.8
3	15	0,15	0.95
4	5	0.05	1
Totale complessivo	100	1	



Esempio: Marada Inn

Guests staying at Marada Inn were asked to rate the quality of their accommodations as being ***excellent. above average. average. below average. or poor.***

The ratings provided by a sample of 20 guests are shown below.

Below Average	Average	Above Average
Above Average	Above Average	Above Average
Above Average	Below Average	Below Average
Above Average	Excellent	Above Average
Average	Above Average	Average
Above Average	Average	Average
Poor	Poor	

Esempio: Marada Inn

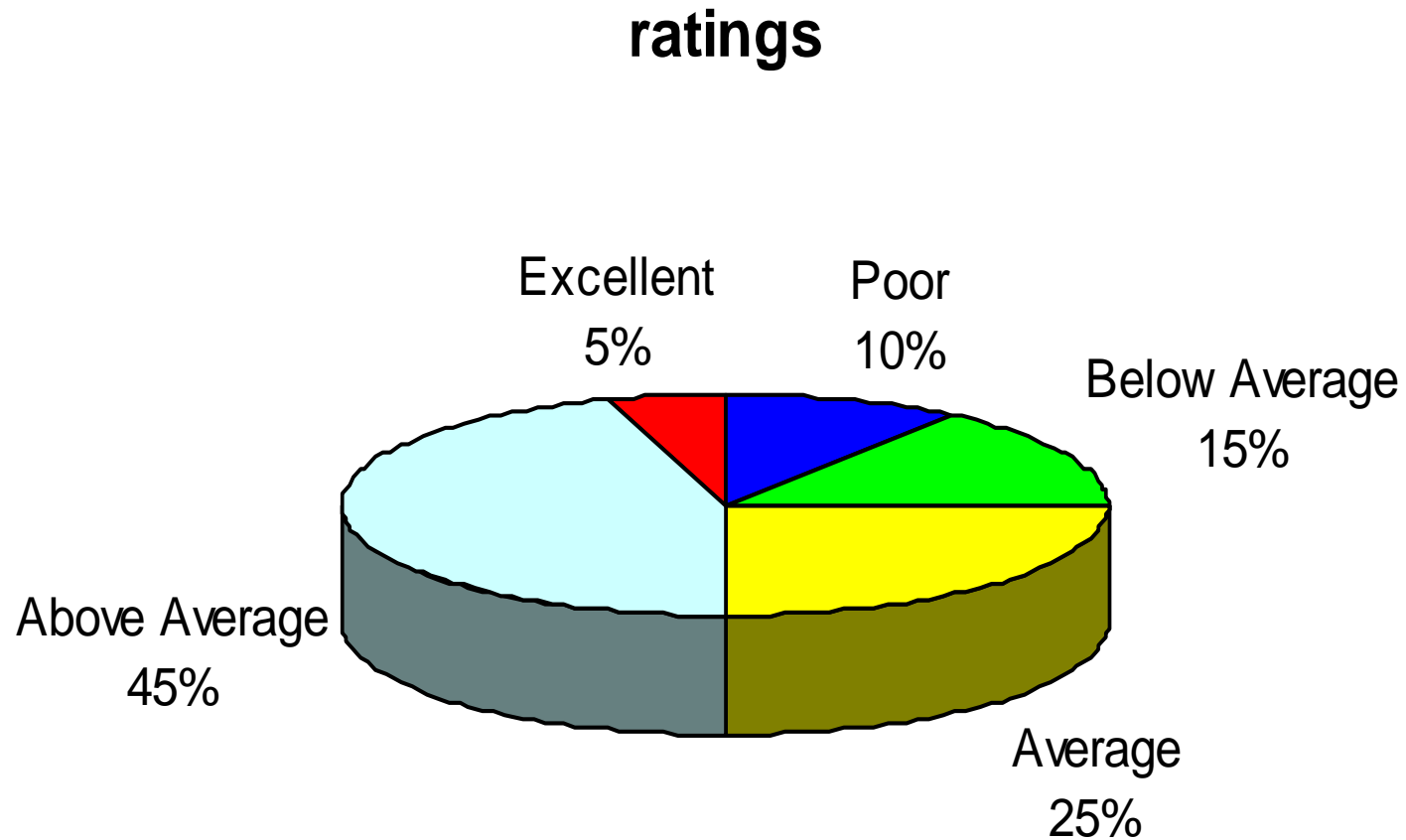
Distribuzione di frequenza

Rating	Frequency
Poor	2
Below Average	3
Average	5
Above Average	9
Excellent	1
Total	20

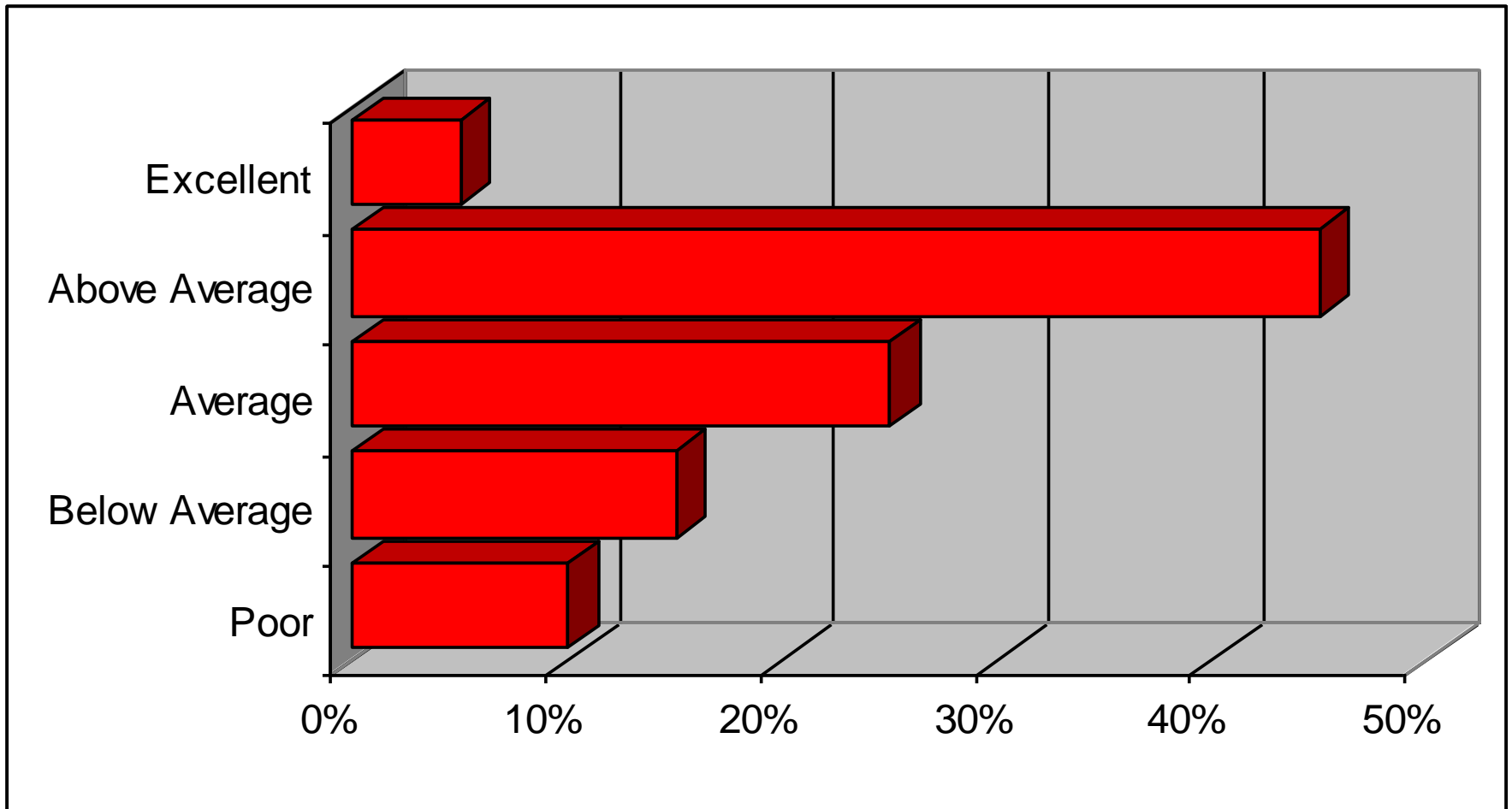
Distribuzione di frequenza: frequenze relative e percentuali

Rating	Relative Frequency	Percent Frequency
Poor	0.10	10%
Below Average	0.15	15%
Average	0.25	25%
Above Average	0.45	45%
Excellent	0.05	5%
Total	1.00	100%

Example: Marada Inn: Pie Chart



Example: Marada Inn: Bar Graph



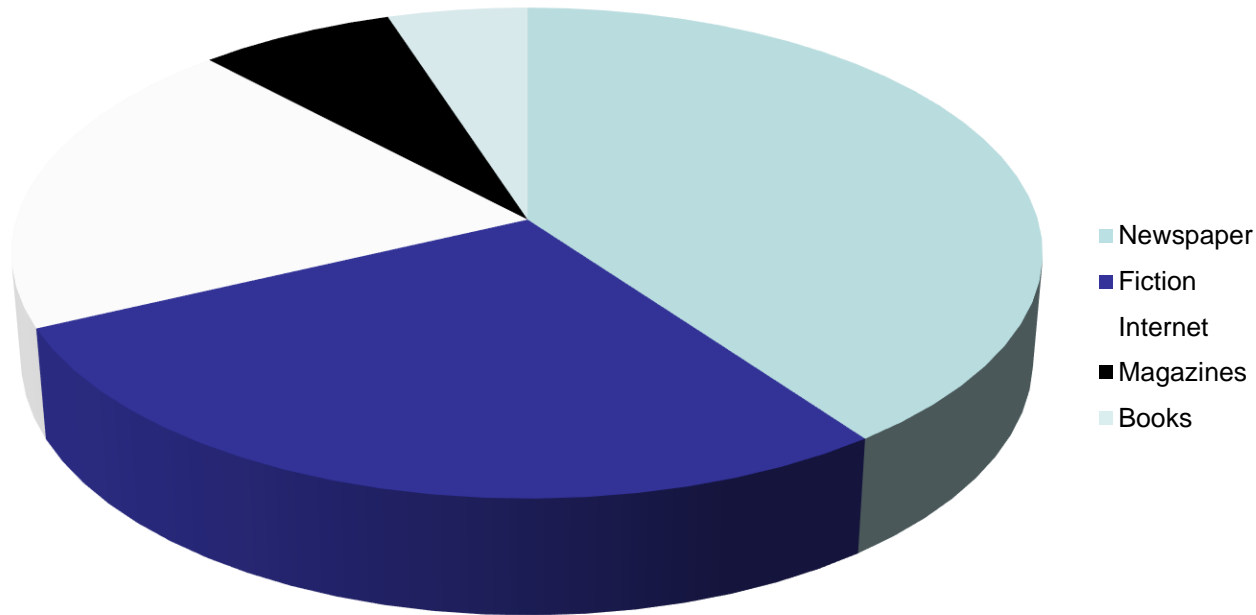
Example: Marada Inn

Insights Gained from the Preceding Pie Chart

- One-half of the customers surveyed gave Marada a quality rating of “above average” or “excellent” (looking at the left side of the pie). This might please the manager.
- For each customer who gave an “excellent” rating, there were two customers who gave a “poor” rating (looking at the top of the pie). This should displease the manager.

Exercise: Interpreting a pie

- The pie chart shows most frequent reading every day



Exercise: Interpreting a pie

- Are people more likely to read fiction or magazines?
- Which two types of reading are more popular than internet?
- Which type of reading are the less common one?

Summarizing Quantitative Data

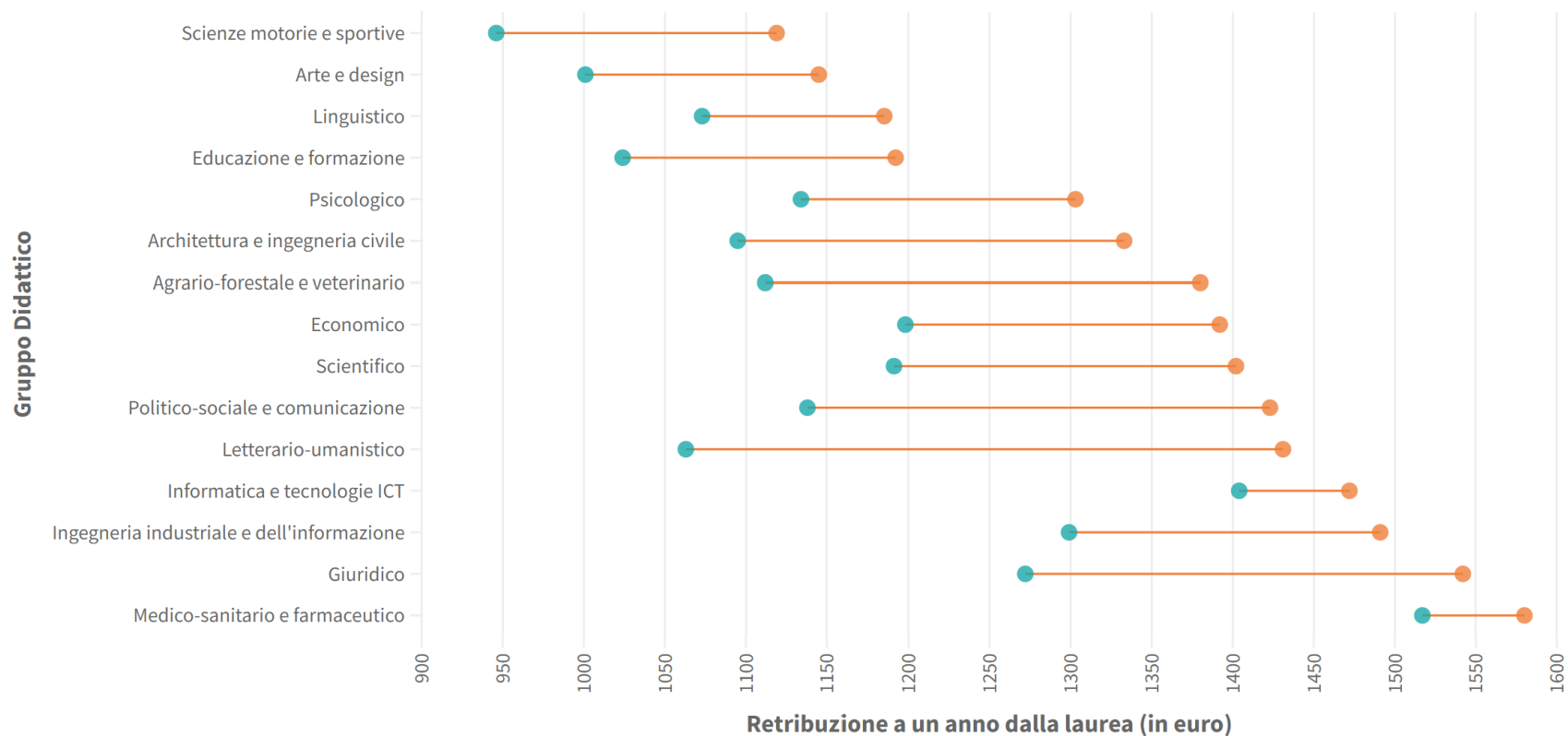
- Frequency Distribution
- Relative Frequency and Percent Frequency
- Histogram
- Cumulative Distributions

Stipendi medi a un anno dalla laurea triennale

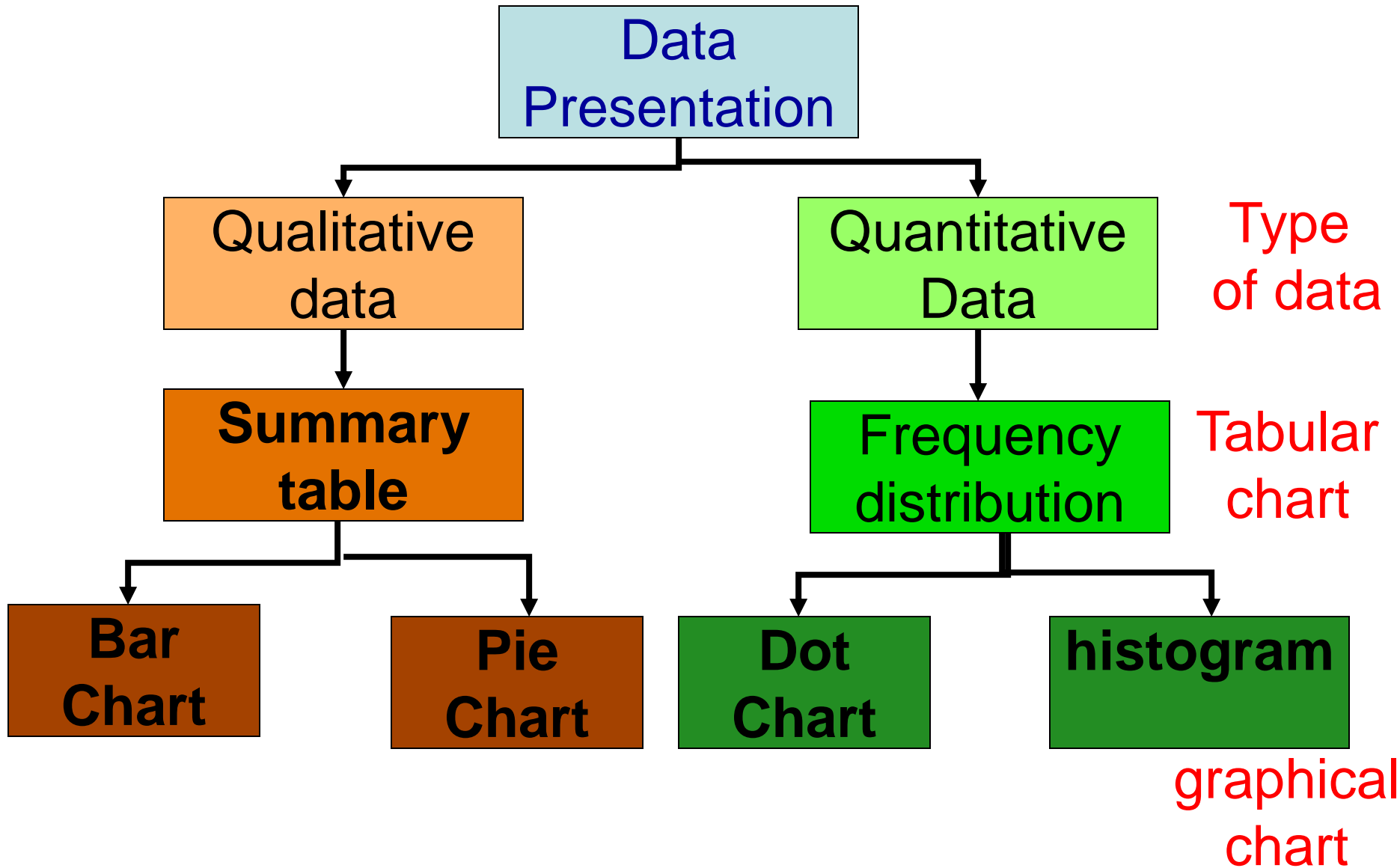
Stipendi medi a un anno dalla laurea di secondo livello

Atenei e Gender pay gap

uomo donna



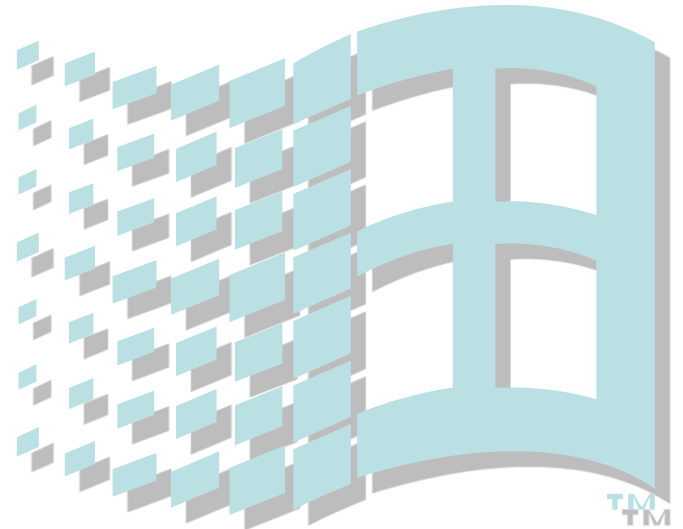
Data Presentation



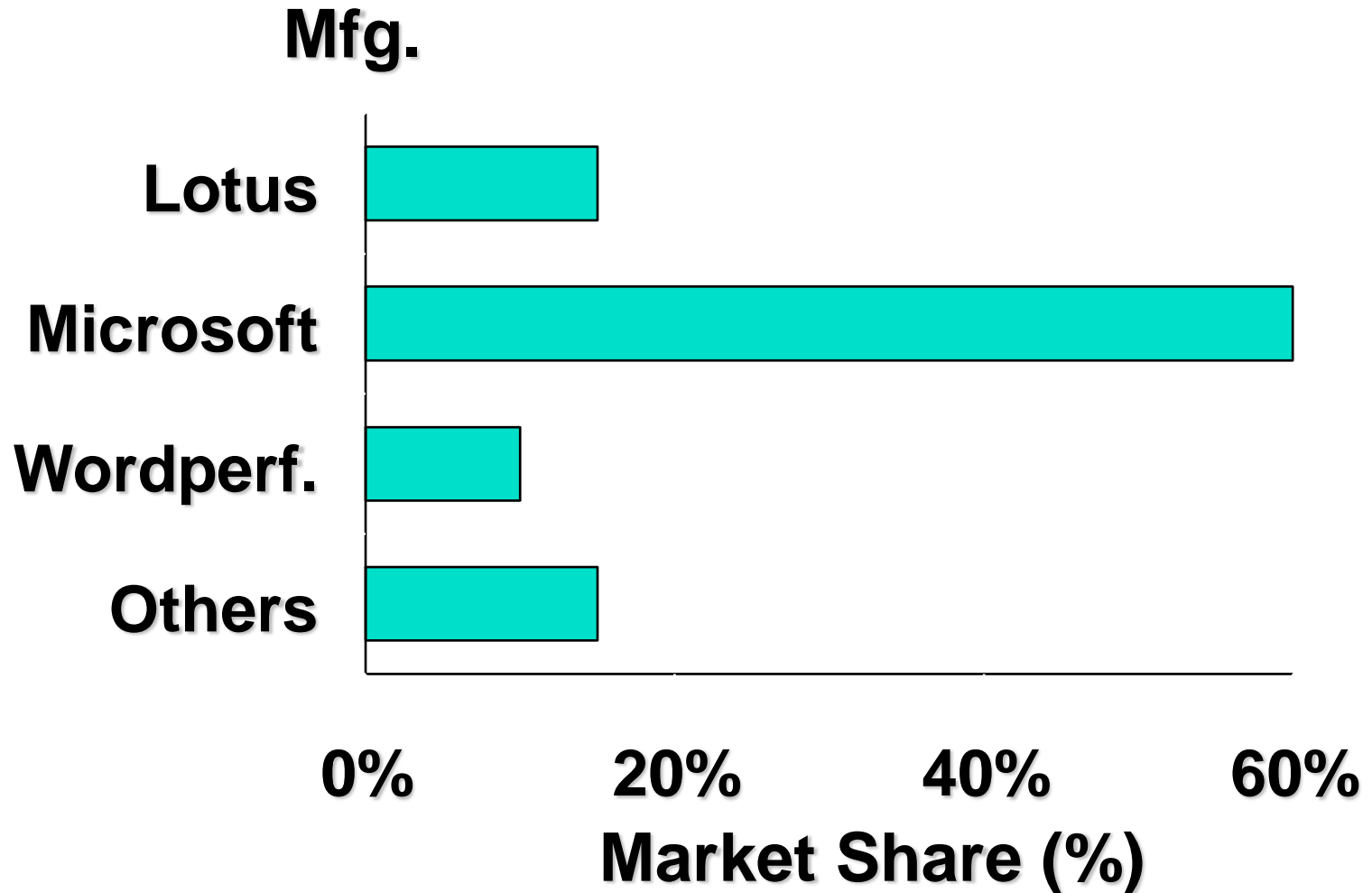
Thinking Challenge

• You're an analyst for IRI.
You want to show the
market shares held by
Windows program
manufacturers in 1992.
Construct a **bar** chart. **pie**
chart.

• Mfg.	Mkt. Share (%)
• Lotus	15
• Microsoft	60
• WordPerfect	10
• Others	15

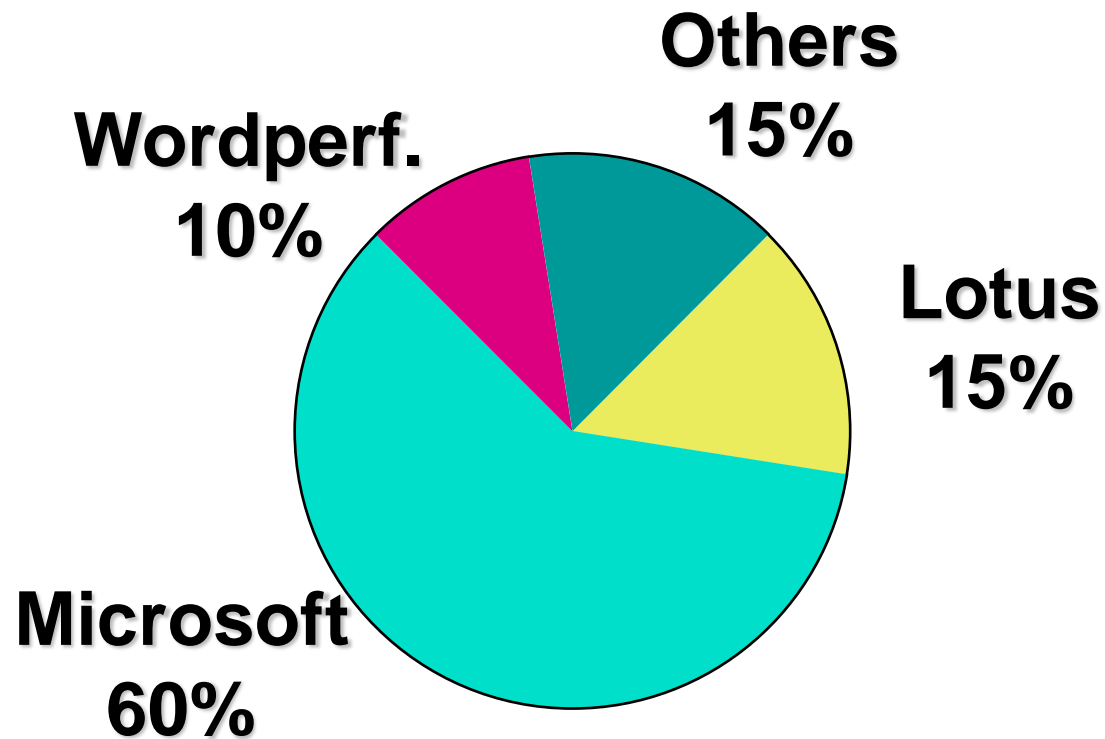


Bar Chart Solution



Pie Chart Solution

Market Share



La suddivisione in classi

Arbitrarietà della suddivisione in classi.

Linee guida:

- ▶ al fine di facilitare l'interpretazione della distribuzione, qualora possibile, le classi dovrebbero avere la stessa ampiezza
- ▶ evitare di costruire classi caratterizzate da un numero di frequenze molto basso
- ▶ equilibrio tra due esigenze in conflitto: sintesi e grado di risoluzione

Attenzione: la suddivisione in classi comporta una perdita di informazioni (le differenze presenti entro la classe).

Tale operazione ha senso soltanto se l'obiettivo finale è produrre una tabella di sintesi o un istogramma.

Per tutti gli altri scopi occorre lavorare con la distribuzione unitaria di partenza.

Rappresentazione grafica della distribuzione di frequenza per caratteri quantitativi continui

Supponiamo di aver ripartito le modalità in classi.

Se le classi sono di ampiezza diversa, le frequenze non sono direttamente confrontabili.

Definiamo allora la *densità di frequenza* (indicata con h_j) come il rapporto tra la frequenza (assoluta o relativa) e l'ampiezza (indicata con a_j) di una classe:

$$h_j = \frac{n_j}{a_j}; \quad \text{ovvero} \quad h_j = \frac{f_j}{a_j}.$$

Istogramma di frequenza

Ad ogni classe è associato un rettangolo, tale che:

- ▶ la base è pari a a_j
- ▶ l'altezza è pari a h_j

L'area del rettangolo è dunque pari alla frequenza (assoluta $n_j = a_j \times h_j$ o relativa $f_j = a_j \times h_j$) associata alla classe.

L'area totale è pari a n o 1.

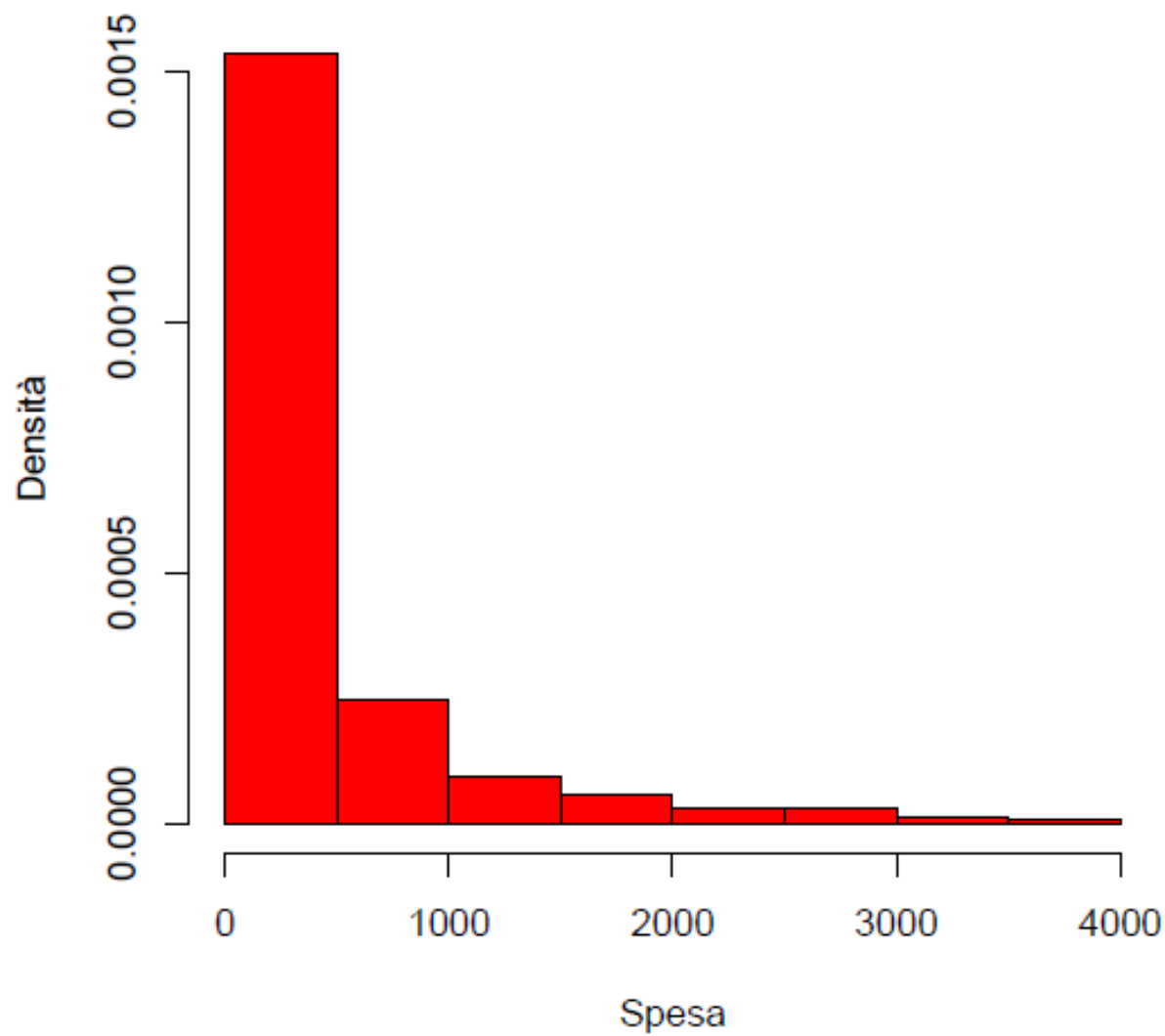


Figura: Istogramma di frequenze relative: classi di uguale ampiezza (500)

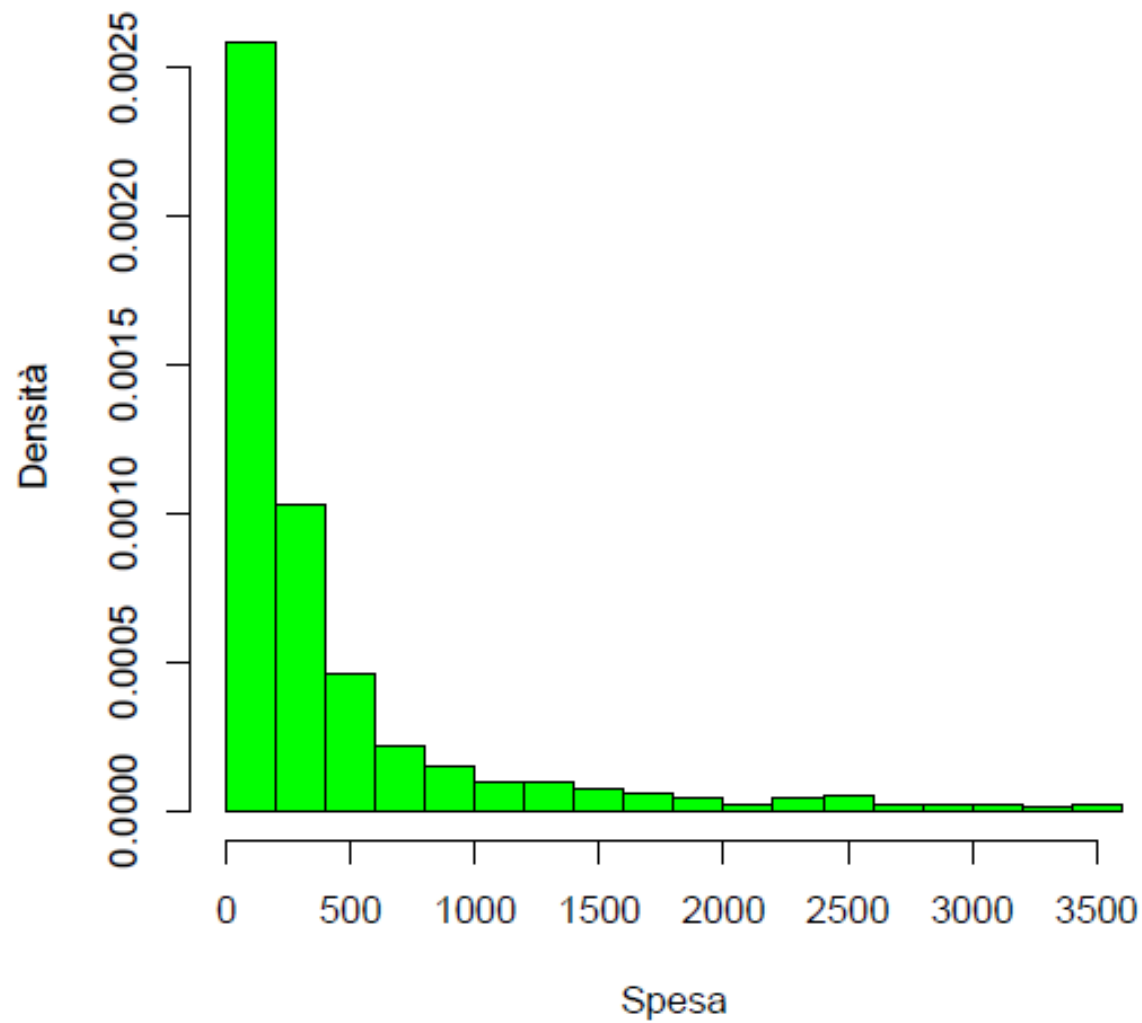


Figura: Istogramma di frequenze relative: classi di uguale ampiezza (200)

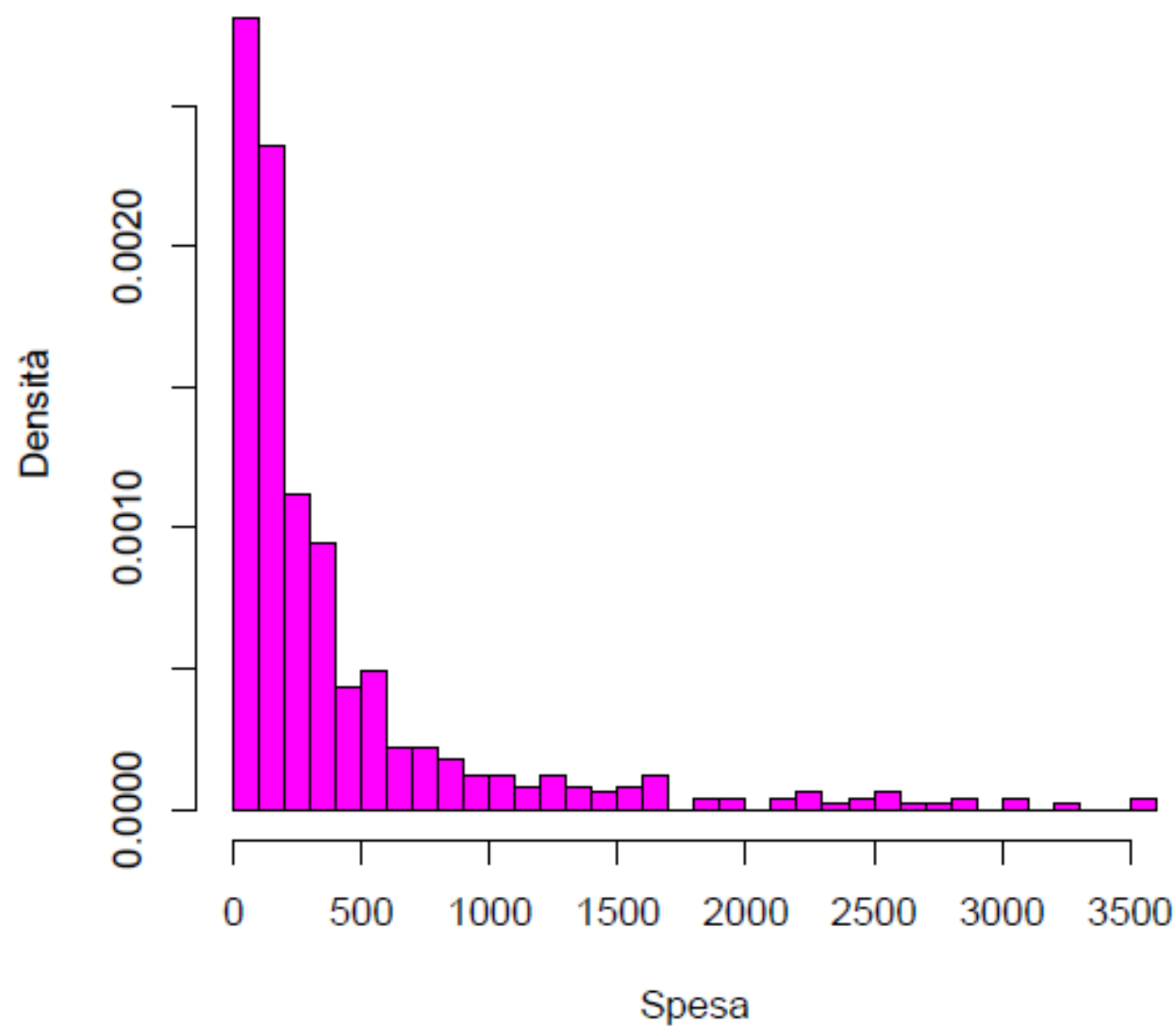


Figura: Istogramma di frequenze relative: classi di uguale ampiezza (100)

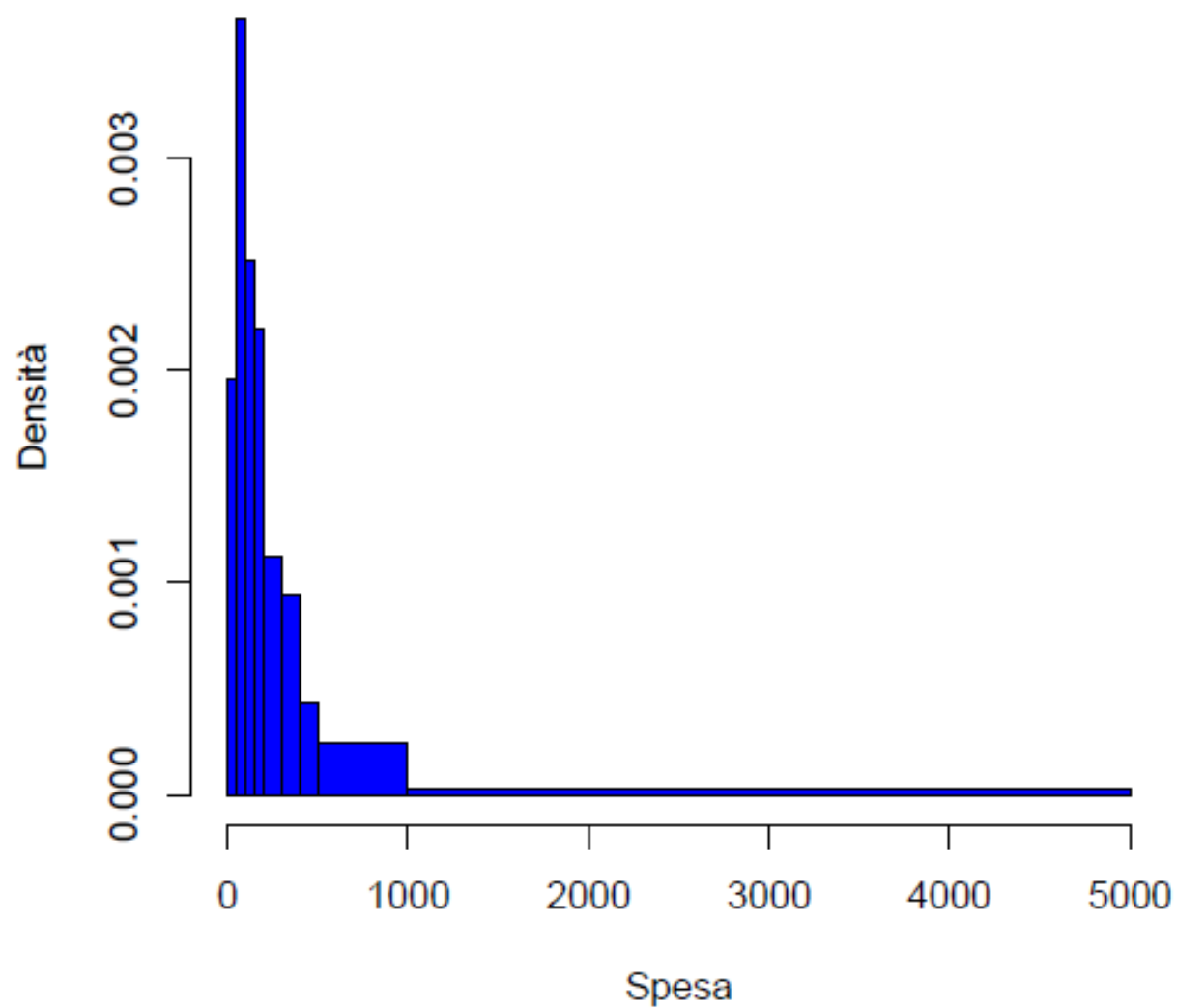
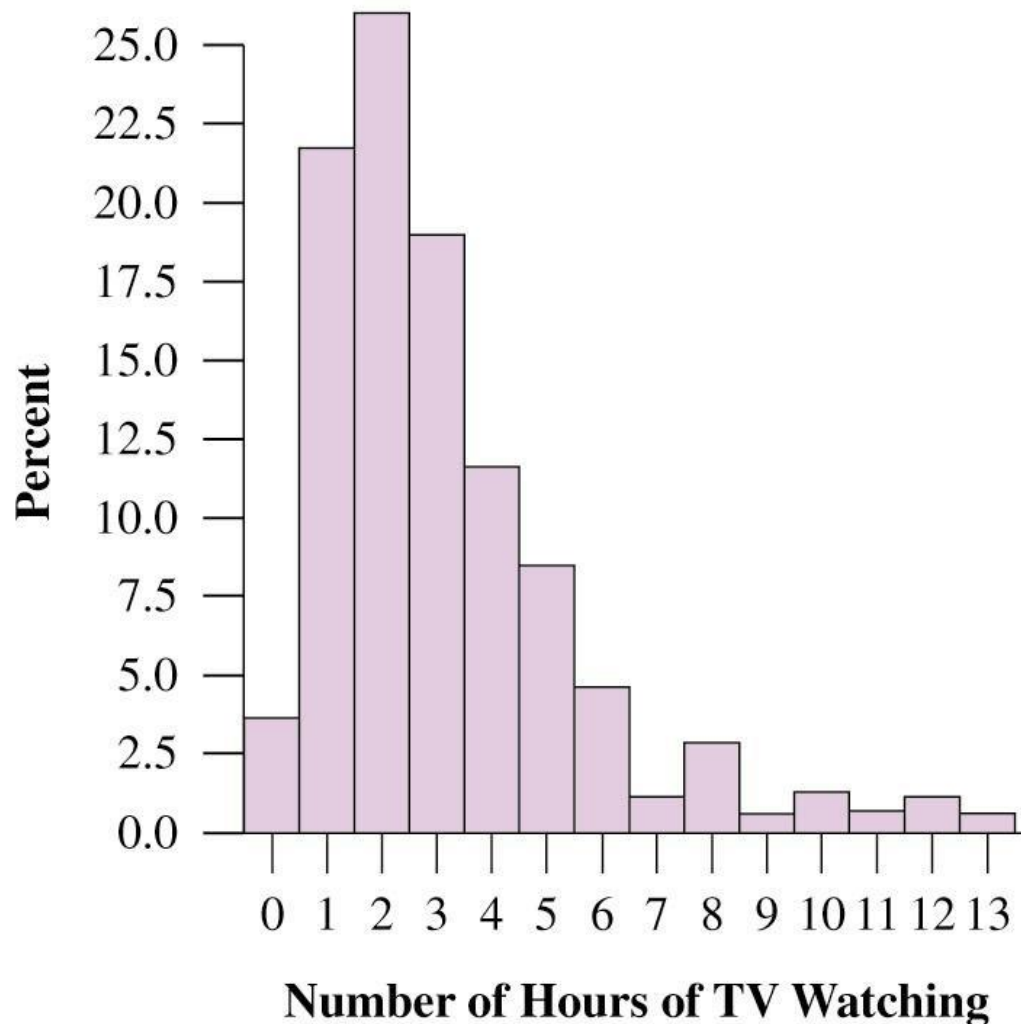
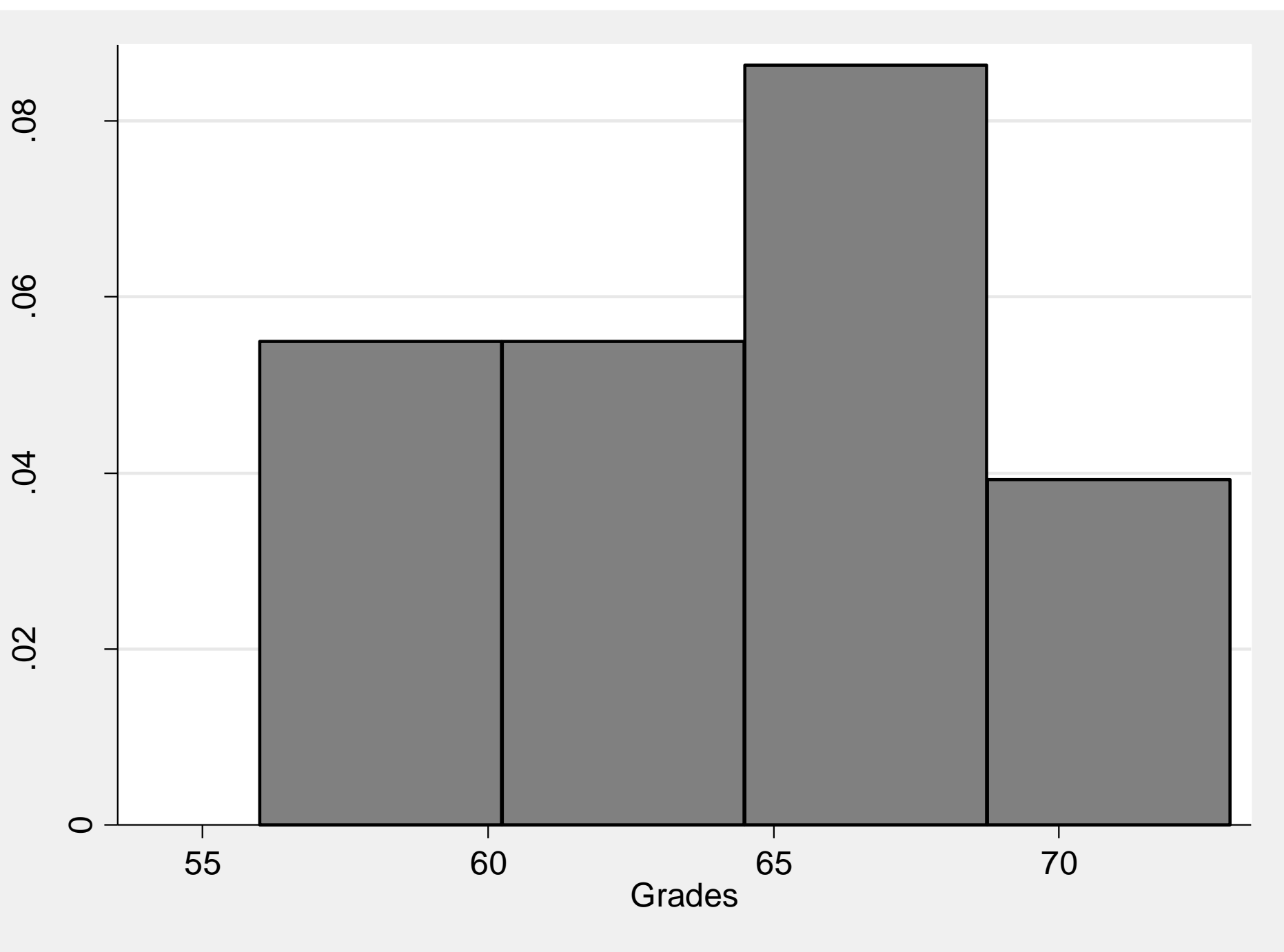
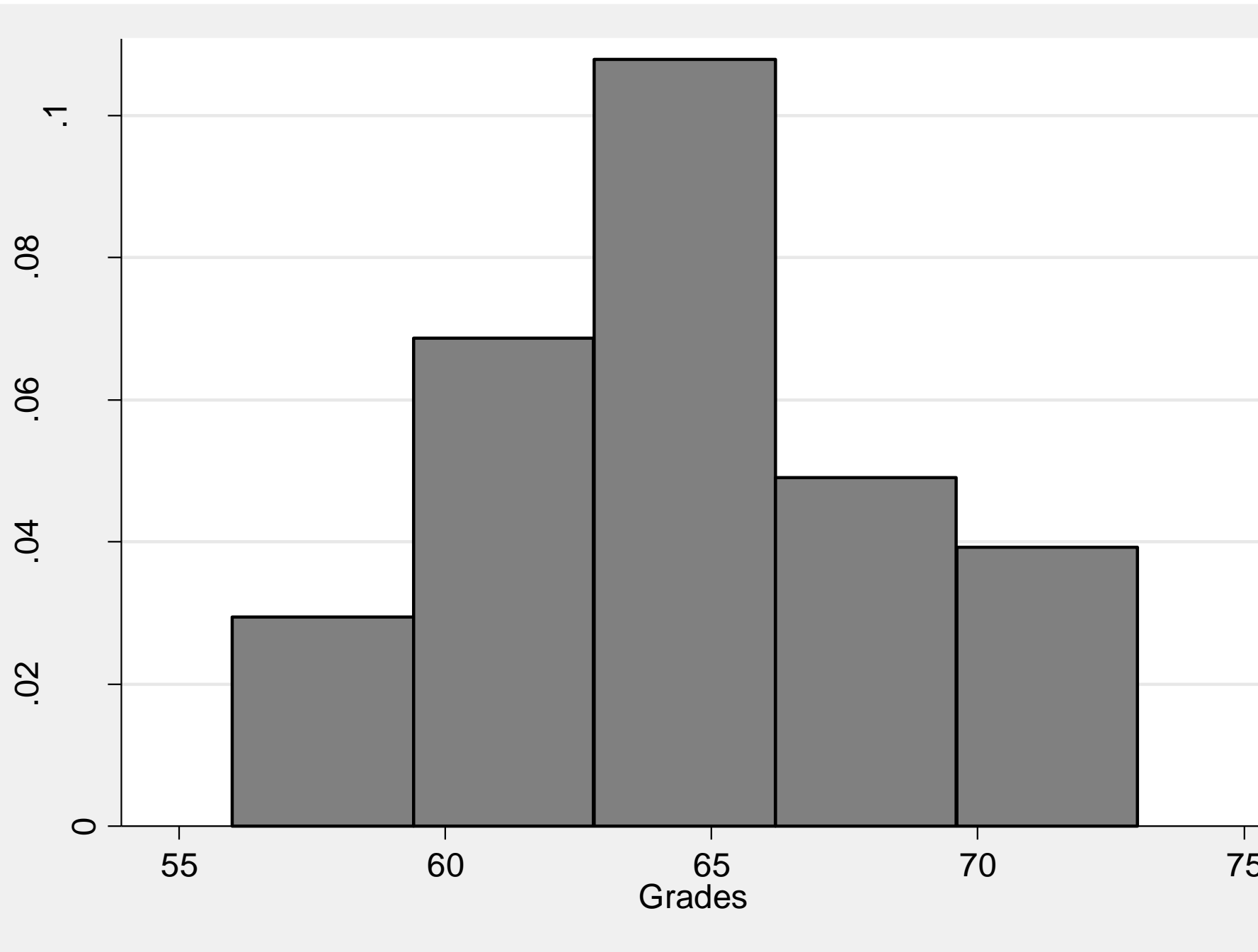


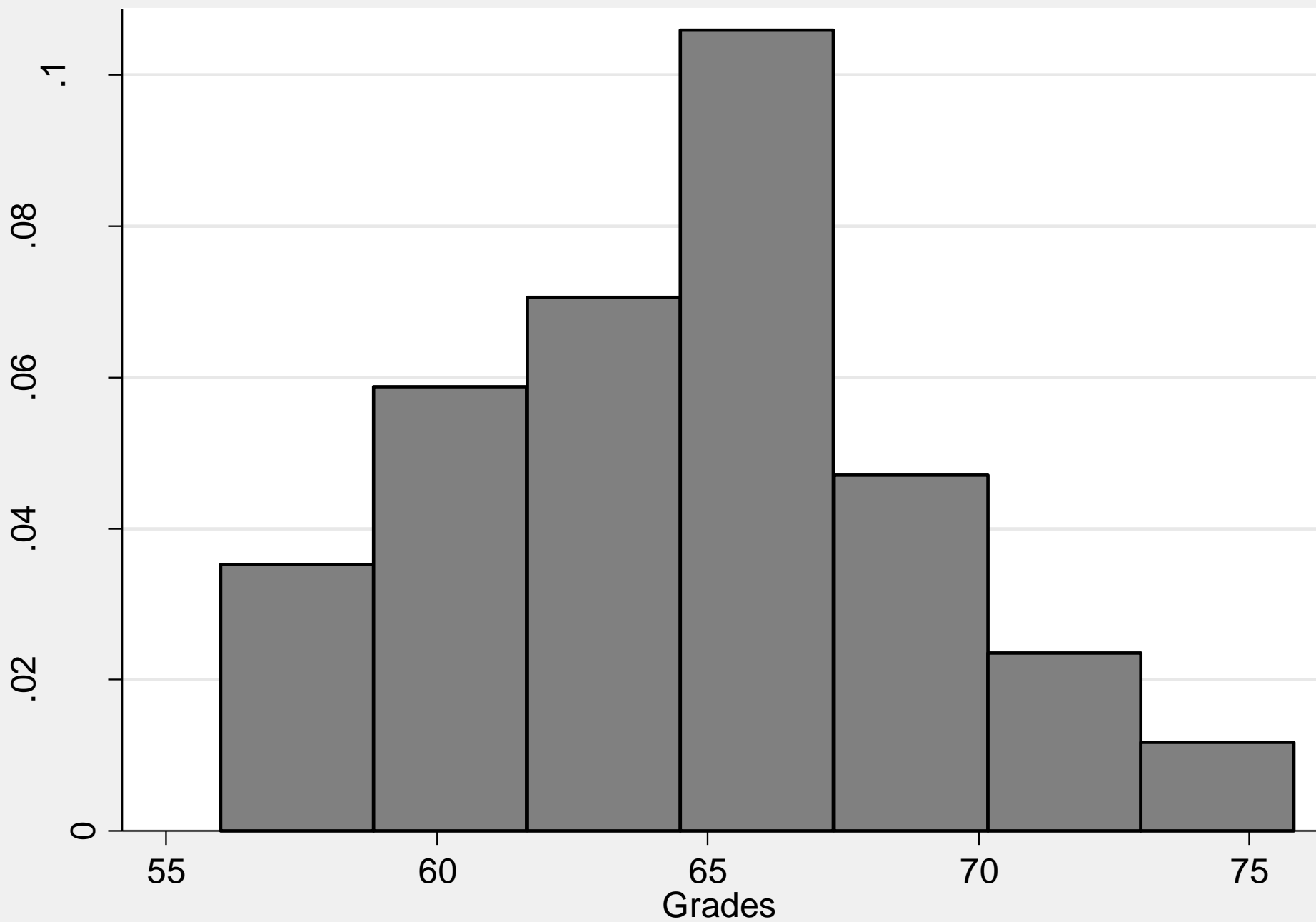
Figura: Istogramma di frequenze relative: classi di ampiezza diversa

- A Histogram is a graph that uses bars to portray the frequencies or the relative frequencies of the possible outcomes for a quantitative variable









Example: Hudson Auto Repair

The manager of Hudson Auto would like to have a better understanding of the cost of parts used in the engine tune-ups performed in the shop. She examines 50 customer invoices for tune-ups. The costs of parts, rounded to the nearest dollar, are listed on the next slide.



Example: Hudson Auto Repair



Parts Cost (\$) for 50 Tune-ups

91	78	93	57	75	52	99	80	97	62
71	69	72	89	66	75	79	75	72	76
104	74	62	68	97	105	77	65	80	109
85	97	88	68	83	68	71	69	67	74
62	82	98	101	79	105	79	69	62	73

Example: Hudson Auto Repair

Cost (\$)	Frequency	Cumulative Frequency	Cumulative Relative Frequency	Cumulative Percent Frequency
[50, 60)	2	2	.04	4
[60,70)	13	15	.30	30
[70,80)	16	31	.62	62
[80,90)	7	38	.76	76
[90,100)	7	45	.90	90
[100, 110]	5	50	1.00	100

$2 + 13$

$15/50$



Example: Hudson Auto Repair

Cost (\$)	Frequency	Relative Frequency	Cumulative Relative Frequency	Cumulative Percent Frequency
[50, 60)	2	0.04	.04	4
[60,70)	13	0.26	.30	30
[70,80)	16	0.32	.62	62
[80,90)	7	0.14	.76	76
[90,100)	7	0.14	.90	90
[100, 110]	5	0.10	1.00	100

Callout for Relative Frequency of [80,90): $13/50$

Callout for Cumulative Relative Frequency of [80,90): $.30 + .32$



Tabular Summary: Frequency and Percent Frequency

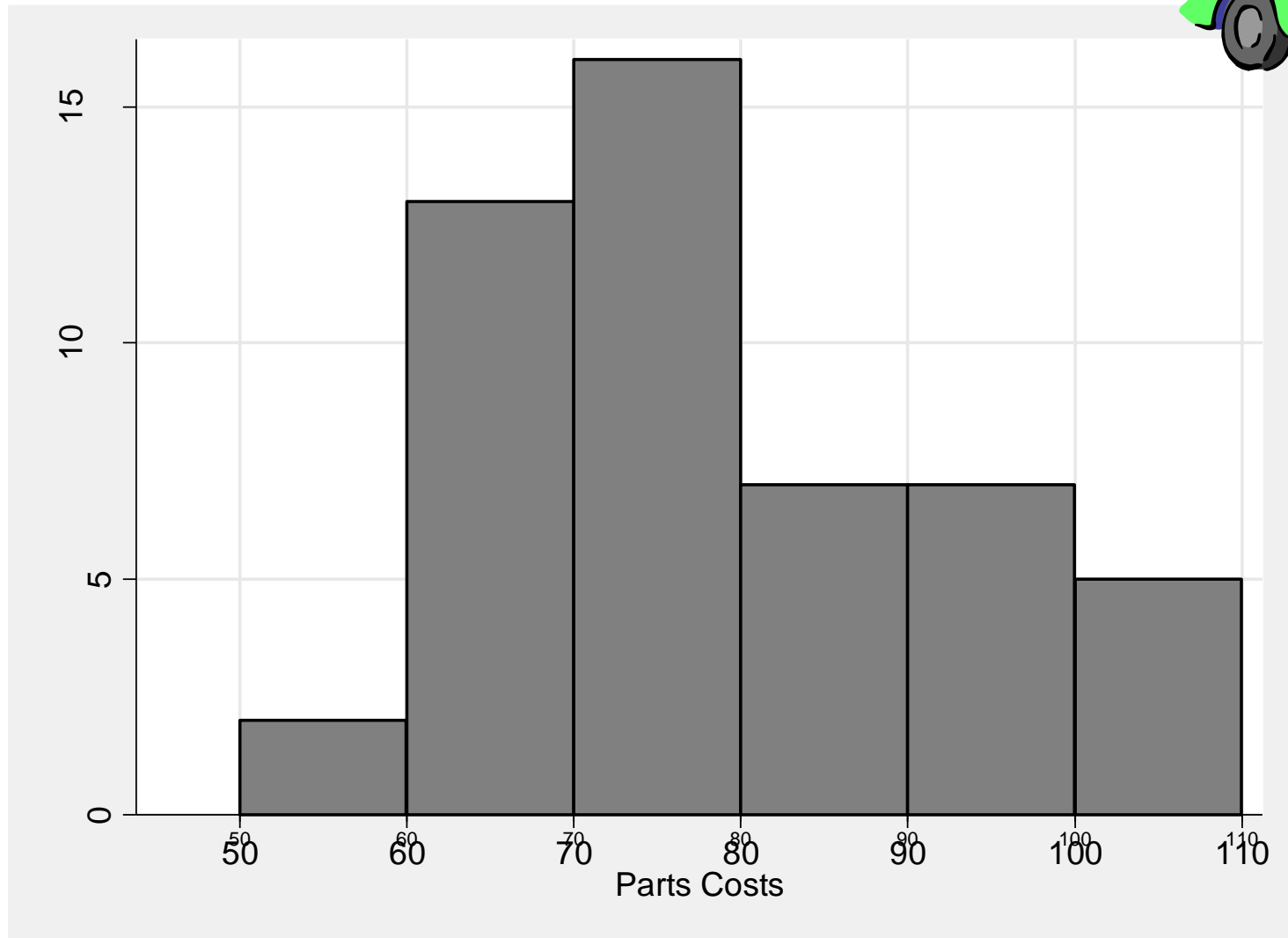


Parts Cost (\$)	Parts Frequency	Percent Frequency
[50,60)	2	4
[60,70)	13	26
[70,80)	16	32
[80,90)	7	14
[90,100)	7	14
[100,110]	<u>5</u>	<u>10</u>
	50	100



$(2/50)100$

Histogram



Frequency Distribution

Guidelines for Selecting Number of Classes

- Use between 5 and 20 classes.
- Data sets with a larger number of elements usually require a larger number of classes.
- Smaller data sets usually require fewer classes.

Frequency Distribution

Guidelines for Selecting Width of Classes

- Use classes of equal width.
- Approximate Class Width =

$$\frac{\text{Largest Data Value} - \text{Smallest Data Value}}{\text{Number of Classes}}$$

Example: Hudson Auto Repair: Frequency distribution

If we choose six classes: approximate Class Width =
 $(109 - 52)/6 = 9.5 \cong 10$

<u>Cost (\$)</u>	<u>Frequency</u>
[50,60)	2
[60,70)	13
[70,80)	16
[80,90)	7
[90,100)	7
[100,110]	5
Total	50

Example: Hudson Auto Repair

Cost (\$)	Relative Frequency	Percent Frequency
[50, 60)	0.04	4
[60, 70)	0.26	26
[70, 80)	0.32	32
[80, 90)	0.14	14
[90, 100)	0.14	14
[100, 110]	0.10	10
Total	1.00	100

Histogram: classes different width

$$\text{density} = \frac{\text{relative frequency}}{\text{class width}}$$

Histogram: classes with different width

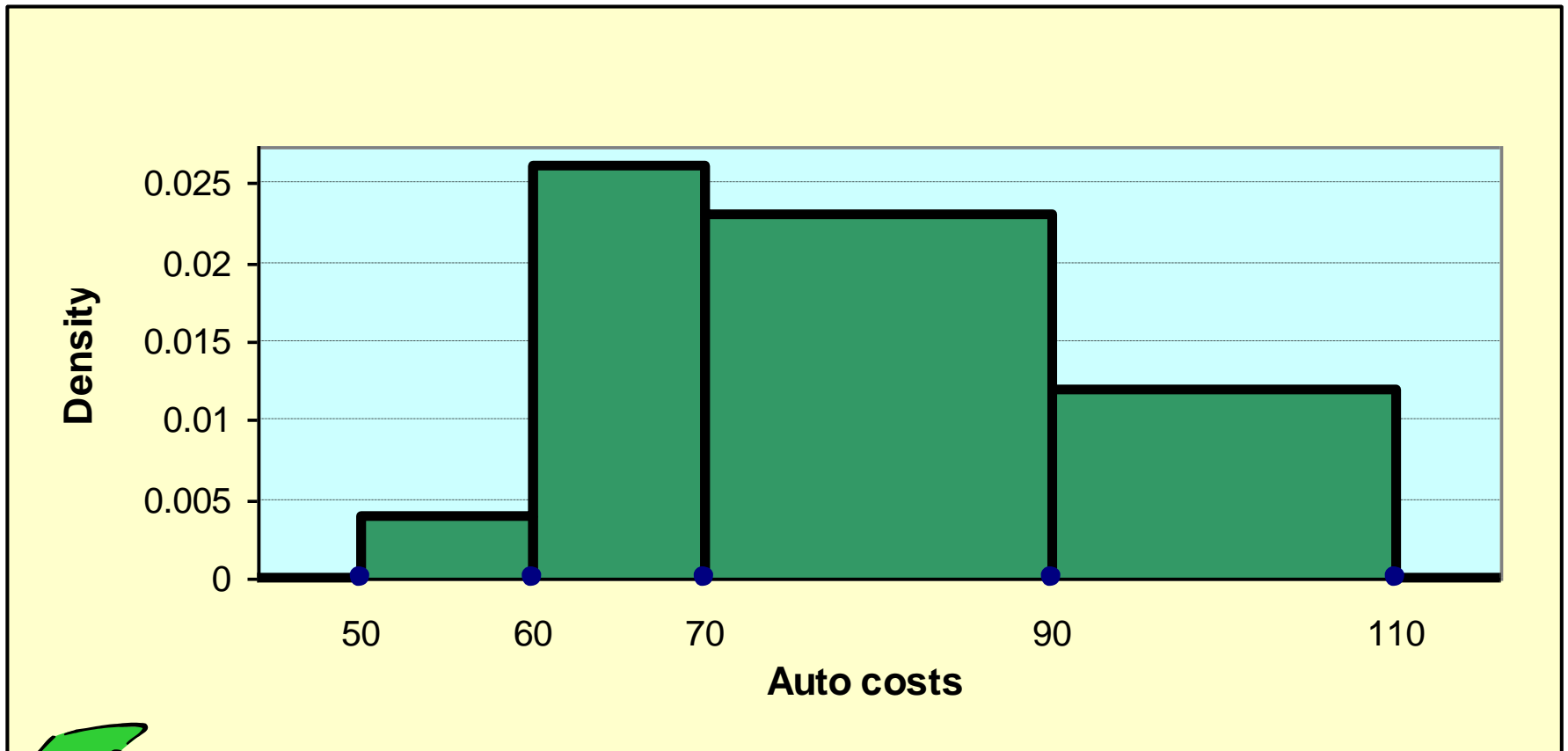
<i>Class</i>	<i>Class width</i>	<i>freq</i>	<i>relative frequency</i>	<i>density</i>
[50, 60)	10	2	0.04	0.004
[60, 70)	10	13	0.26	0.026
[70, 90)	20	23	0.46	0.023
[90, 110]	20	12	0.24	0.012



0.04/10

0.24/20

Histogram: classes with different width



Histogram: classes with different width

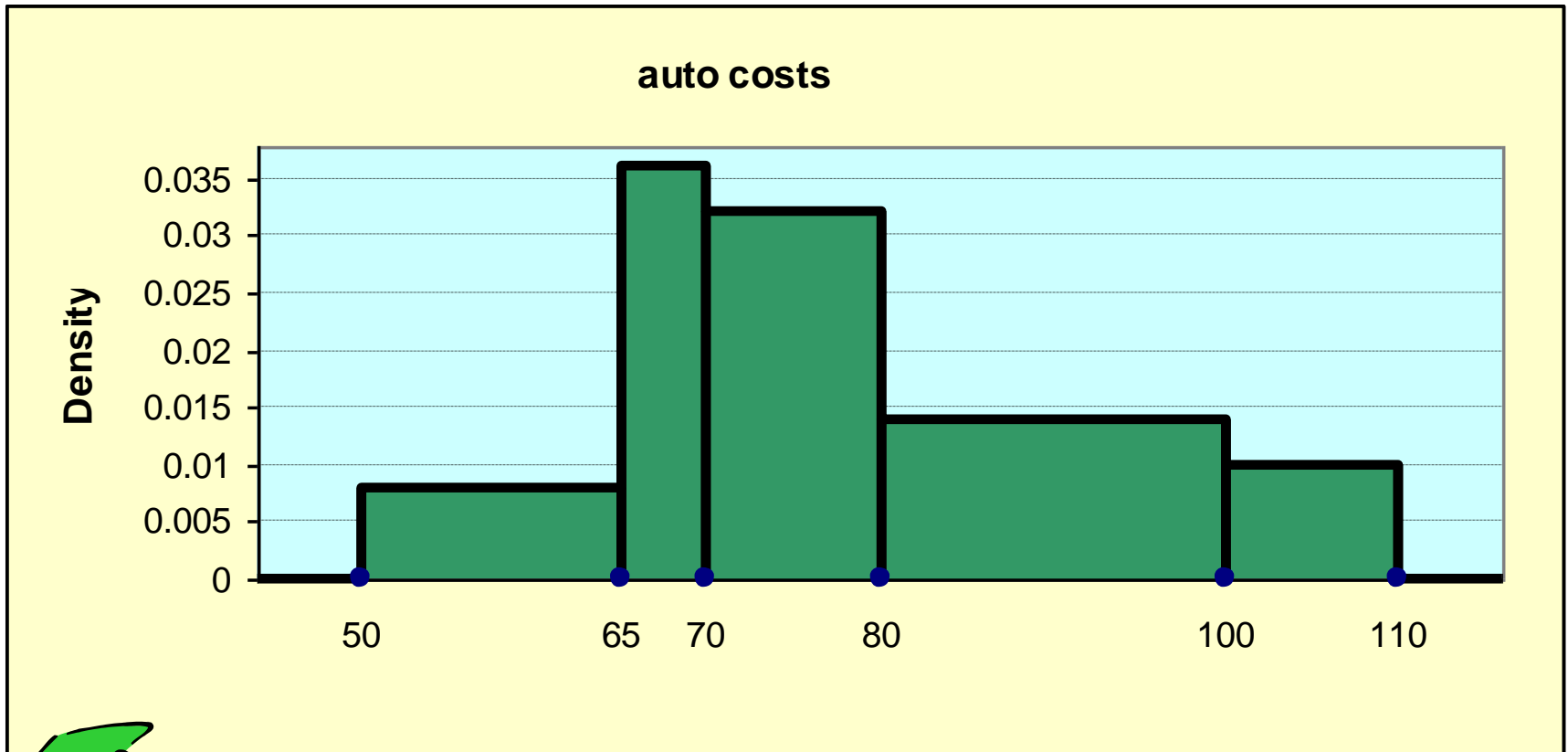
Class	Class width	Freq	relative frequency	Density
[50, 65)	15	6	0.12	0.008
[65, 70)	5	9	0.18	0.036
[70, 80)	10	16	0.32	0.032
[80,100)	20	14	0.28	0.014
[100, 110]	10	5	0.1	0.01



$0.18/5$

$0.1/10$

Histogram: classes with different width



Example: Hudson Auto Repair

Insights Gained from the Percent Frequency Distribution

- Only 4% of the parts costs are in the \$50-59 class.
- 30% of the parts costs are under \$70.
- The greatest percentage (32% or almost one-third) of the parts costs are in the \$70-79 class.
- 10% of the parts costs are \$100 or more.

