UNIVERSITÀ
DEGLI STUDI DI TRIESTE

Dipartimento di scienze economiche,
aziendali, matematiche e statistiche
“Bruno de Finetti”

Statistica

Distribuzioni statistiche e loro rappresentazioni

Francesco Pauli

A.A. 2017/2018 (aggiornamento 20/2/2020)

Indice

Indice

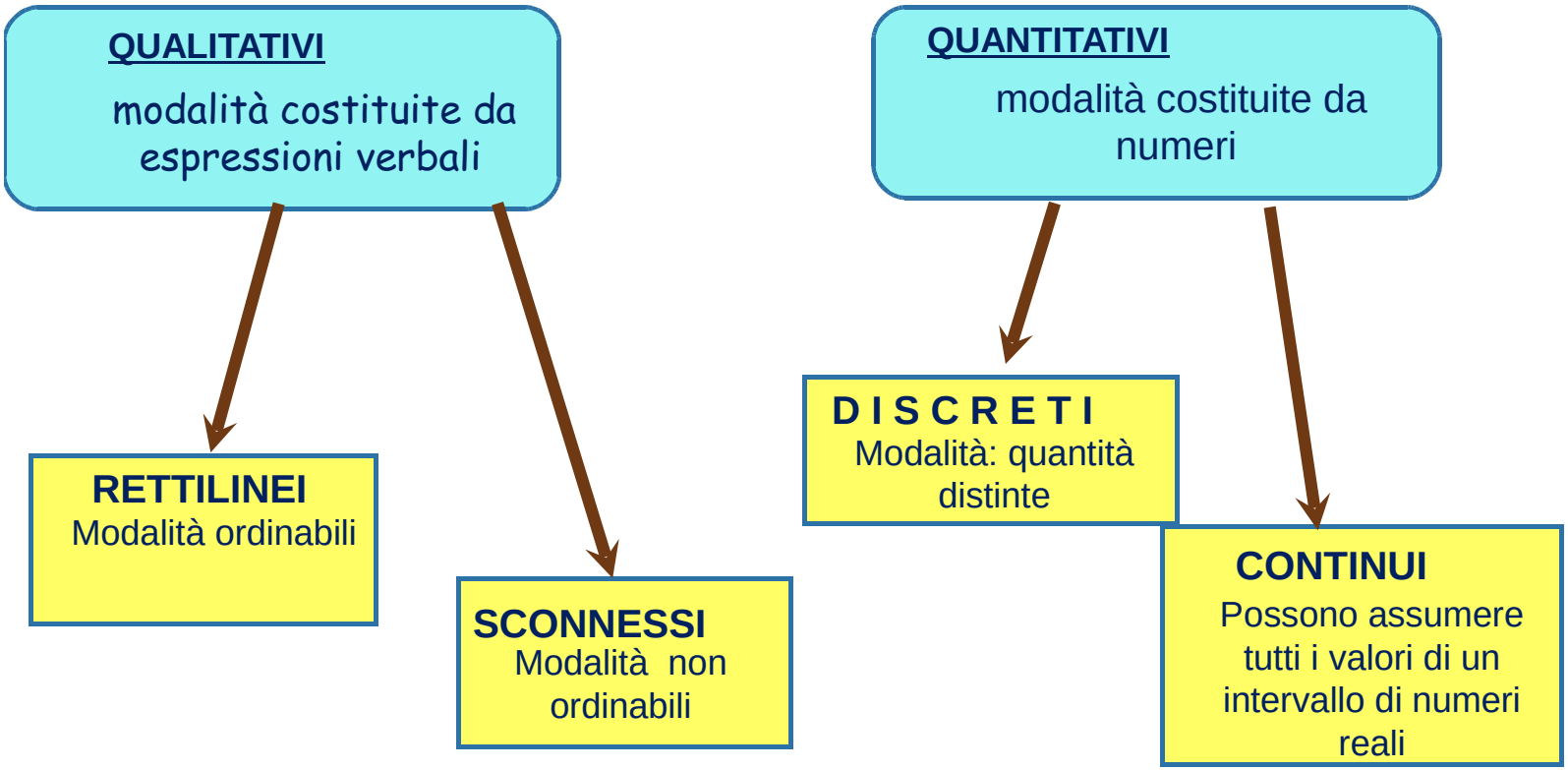
Terminologia elementare

Un dato statistico è il risultato della rilevazione (misurazione/osservazione) di un qualche **carattere** su un' **unità statistica** appartenente a una popolazione.

- **unità statistica**: il caso individuale componente del collettivo statistico;
- **carattere** (o **variabile**): ogni aspetto elementare oggetto di rilevazione nelle unità statistiche del collettivo;
- **modalità** di un carattere: i diversi modi con cui questo si presenta nelle unità statistiche del collettivo
- **supporto**: insieme (teorico) delle modalità di un carattere.

Nel seguito, i termini **carattere** e **variabile** verranno usati in modo interscambiabile.

Tipi di carattere



Variabili qualitative

- Una variabile è **qualitativa** se le modalità che si presentano sono espresse in forma verbale;
  - una variabile qualitativa è **sconnessa** se le sue modalità non implicano una graduazione (mutabile sconnessa);
  - una variabile qualitativa è **ordinale** se le sue modalità implicano una graduazione;
- le modalità possono essere predefinite a priori;
- a volte, nelle indagini, le modalità vengono desunte a posteriori dalla descrizione dettagliata che il rilevatore fa dello stato della singola unità relativamente al carattere in questione.

Esempio: qualitativa sconnessa

Ti è piaciuta l'ultima edizione del Festival di Sanremo?

- L'ho visto e mi è piaciuto
- L'ho visto e non mi è piaciuto
- Non l'ho visto

Esempio: qualitativa sconnessa 2

Qual è il tuo genere letterario preferito?

- Comico/umoristico
- Fantascienza
- Fantasy
- Giallo/noir/thriller
- Psicologico
- Romantico
- Storico
- Altro

Esempio: qualitativa ordinale

Quanto frequentemente bevi birra?

- Mai
- Una volta a settimana
- Più volte a settimana
- Ogni giorno
- Più volte al giorno

Variabili quantitative

Una variabile è **quantitativa** se le modalità che si presentano sono espresse in forma numerica.

Distinguiamo rispetto

- ai valori che possono assumere
  - una variabile quantitativa è **discreta** se l'insieme delle sue modalità è finito oppure numerabile (detto in altri termini, se la quantità che rappresenta varia "a salti");
  - una variabile quantitativa è **continua** se l'insieme delle sue modalità è un intervallo, limitato o illimitato.

  (Per la limitata precisione utilizzabile nel rilevare le misure, la distinzione tra variabile discreta e continua è convenzionale.)
- e rispetto alle operazioni che è ragionevole fare
  - una variabile è **intervallare** se ha senso fare differenze tra valori ma non c'è uno zero naturale e non ha senso fare rapporti
  - una variabile è **rapportabile** se ha senso fare rapporti tra valori (c'è uno zero naturale)

Esempi: variabili quantitative

- Discreta rapportabile

  Quante volte sei stato al cinema negli ultimi tre mesi?

- Continua rapportabile

  Qual è la tua altezza (in centimetri)?

- Discreta non rapportabile

  In che anno sei nato?

- Continua non rapportabile

  Temperatura esterna

Esercizio

Che tipo di carattere è il colore dei capelli?

(a) numerica, continua
(b) numerica, discreta
(c) qualitativa, sconnessa
(d) qualitativa, ordinale

Esercizio

Che tipo di carattere è il voto di maturità?

(a) numerica, continua
(b) numerica, discreta
(c) qualitativa, sconnessa
(d) qualitativa, ordinale

Esercizio

Che tipo di carattere è il reddito personale?

(a) numerica, continua
(b) numerica, discreta
(c) qualitativa, sconnessa
(d) qualitativa, ordinale

Indice

Matrice dei dati

I dati vengono organizzati in una matrice, ad esempio i dati del questionario (2015/16):

variabile ↓

	sesso	Sanremo	···	sonno	studio
1	maschio	Non l'ho visto	···	8	2
2	femmina	L'ho visto e mi è piaciuto	···	6	30
3	maschio	Non l'ho visto	···	9	5
4	femmina	Non l'ho visto	···	8	25
⋮	⋮	⋮	⋮	⋮	⋮
	femmina	Non l'ho visto	···	8	20

← unità statistica

Genere	AnnoCorso	Residenza	Altezza	Peso	OreStudioSett	OreSonnoNotte	VotoMatura	E
Maschio	Secondo	Italia	178.00	60.00	18.00	8.00	73.00	N
Maschio	Secondo	TS	180.00	90.00	10.00	7.00	73.00	S
Maschio	Secondo	Regione limitrofa	175.00	80.00	30.00	9.00	64.00	S
Maschio	Secondo	FVG	186.00	71.00	30.00	8.00	81.00	S
Maschio	Terzo	FVG	170.00	61.00	30.00	8.00	75.00	N
Femmina	Secondo	FVG	164.00	53.00	20.00	6.00	70.00	S
Femmina	Secondo	TS	158.00	60.00	30.00	5.00	72.00	S
Femmina	Secondo	TS	165.00	48.00	25.00	7.00	73.00	N
Maschio	Secondo	TS	182.00	70.00	24.00	7.00	60.00	S
Maschio	Secondo	TS	172.00	63.00	30.00	7.00	71.00	S
Femmina	Secondo	FVG	158.00	43.00	37.00	7.00	79.00	S
Femmina	Terzo	TS	158.00	47.00	30.00	8.00	74.00	S
Maschio	Secondo	FVG	185.00	73.00	10.00	8.00	83.00	S
Maschio	Secondo	Regione limitrofa	194.00	89.00	30.00	8.00	81.00	S
Femmina	Secondo	FVG	170.00	57.00	40.00	8.00	100.00	S
Femmina	Secondo	TS	162.00	70.00	6.00	7.00	77.00	N
Femmina	Secondo	TS	162.00	53.00	20.00	7.00	68.00	S
Maschio	Secondo	Regione limitrofa	169.00	70.00	14.00	7.00	100.00	S
Maschio	Terzo	TS	193.00	85.00	10.00	6.00	94.00	S
Femmina	Secondo	TS				7.00	85.00	S
Femmina	Secondo	FVG	173.00	56.00	4.00	8.00	70.00	S
Maschio	Secondo	TS	186.00	67.00	22.00	8.00	76.00	S
Femmina	Secondo	TS	162.00	52.00	5.00	8.00	82.00	S
Femmina	Secondo	FVG	165.00	67.00	20.00	8.00	90.00	N
Maschio	Secondo	TS	175.00	65.00	20.00	8.00	80.00	S
Femmina	Secondo	TS	163.00	52.00	40.00	7.00	82.00	S
Femmina	Oltre il terzo	TS	160.00	52.00	14.00	8.00	97.00	N
Femmina	Terzo	FVG	170.00	68.00	35.00	6.00	76.00	S
Femmina	Secondo	TS	170.00	56.00		8.00	75.00	S
Maschio	Oltre il terzo	TS	173.00	72.00	40.00	7.00	90.00	S
Maschio	Secondo	FVG	186.00	78.00	10.00	7.00	82.00	S
Maschio	Secondo	TS	188.00	87.00	60.00	8.00	79.00	S
Maschio	Secondo	TS	192.00	70.00	6.00	7.00	84.00	S
Femmina	Terzo	FVG	165.00	44.00	20.00	8.00	76.00	S

L'agente arancio

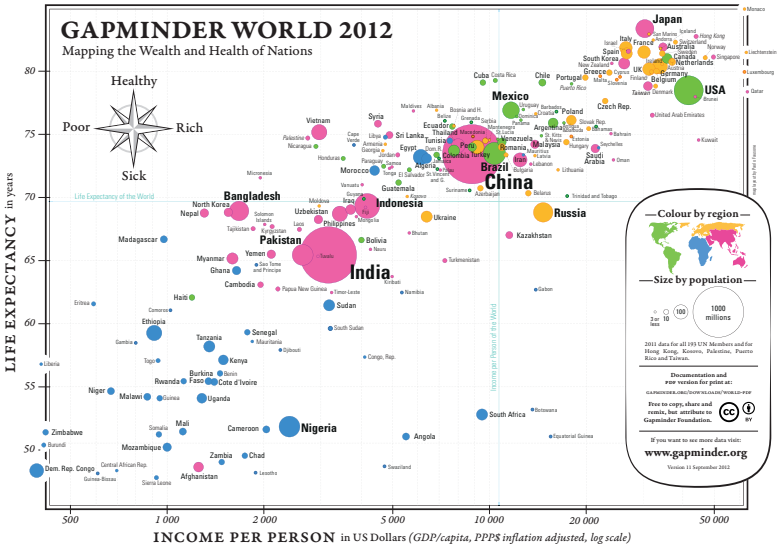
- L'agente arancio è una mistura erbicida ampiamente usata nella guerra del Vietnam.
- L'agente arancio è stato collegato a patologie cancerose in molti studi epidemiologici.
- Per studiare l'assorbimento della diossina, nel 1987 le concentrazioni di diossina (in parti per trilione) vennero misurate nel plasma di veterani (soldati di terra) dell'esercito USA.
- Il campione era così composto
  - campione (non casuale) di veterani del Vietnam che servirono nel 1967-1968
  - campione (non casuale) di veterani che servirono in USA e Germania nel 1965-1971

L'agente arancio (vets): matrice dei dati

SUBJECT	DIOXIN	VETERAN
1	0	VIETNAM
2	0	VIETNAM
3	0	VIETNAM
4	0	VIETNAM
5	0	VIETNAM
.....	....	.....
739	9	OTHER
740	9	OTHER
741	10	OTHER
742	11	OTHER
743	15	OTHER

Dati su salute e ricchezza

I dati per costruire il grafico di gapminder si presentano così



Paese	time	gdppc	lifexp	Regione
Albania	1980	1061	70	Europe & Central Asia
Albania	1990	978	74	Europe & Central Asia
Albania	1991	688	73	Europe & Central Asia
Albania	1992	643	73	Europe & Central Asia
Albania	1993	714	73	Europe & Central Asia
Albania	1994	785	74	Europe & Central Asia
...	...	...	...	...
Zimbabwe	2007	340	48	Sub-Saharan Africa
Zimbabwe	2008	280	49	Sub-Saharan Africa
Zimbabwe	2009	297	50	Sub-Saharan Africa
Zimbabwe	2010	323	50	Sub-Saharan Africa
Zimbabwe	2011	348	52	Sub-Saharan Africa

Dati su famiglie e abitazioni

Qui una parte dei dati della rilevazione Istat sulle famiglie.

	reddito	tel	tv	pc	lavatrice	auto	tipoab	titolo	numstanze
1	18902	si	si	si	si	si	singola	prop	2
2	1684	si	si	no	si	si	singola	prop	4
3	28783	no	no	no	no	si	apppiu10	prop	5
4	17809	si	si	no	si	si	singola	prop	5
5	6976	si	si	si	si	si	appmeno10	gratuito	4
6	6280	si	si	si	si	si	apppiu10	prop	4
7	18088	si	si	no	si	si	apppiu10	prop	3
8	869	no	si	no	si	si	appmeno10	prop	2
9	15017	si	si	no	si	si	...	prop	5
10	25000	si	si	si	si	si	apppiu10	affmerc	4
11	3922	si	si	si	si	si	apppiu10	prop	4
12	72089	si	si	si	si	si	appmeno10	affmerc	6
13	13498	si	si	no	si	si	appmeno10	prop	3
14	7669	si	no	si	si	si	appmeno10	affagev	3
15	34677	si	si	si	si	si	apppiu10	prop	3
16	36939	si	si	si	si	si	apppiu10	affagev	5
17	13639	si	si	si	si	si	apppiu10	affmerc	4
18	28063	si	si	si	si	si	apppiu10	affmerc	4
19	27685	si	si	si	si	si	appmeno10	prop	3
20	1183	si	si	no	si	si	appmeno10	prop	4
21	...	...	...	...	...	...	...	...	...
19142	632	si	si	no	si	si	semisingola	prop	3
19143	7002	si	si	no	si	no	appmeno10	affmerc	3
19144	306	si	si	si	si	si	apppiu10	prop	5
19145	31683	si	si	si	si	si	appmeno10	gratuito	3
19146	910	si	si	no	si	si	appmeno10	prop	4
19147	-897	no	si	no	no	no	...	gratuito	3

Dati su famiglie e abitazioni

Nella stessa rilevazione si considerano anche alcune informazioni individuali

	annonas	sesso	statociv	istruzione	tipolav	tipoctr	reddito	salute	orelav
1	1973	m	lib	secsup	dip	indet	18636	+	30
2	1962	f	lib	secsup	dip	indet	28080	++	50
3	1956	f	lib	secinf	dip	indet	1770	=	...
4	1954	m	con	secinf	autsenzadip	...	15874	=	20
5	1965	f	sep	secsup	autsenzadip	...	0	+	...
6	1990	m	lib	secsup	nonlav	...	0	+	...
7	1968	f	con	secsup	nonlav	...	0	+	...
8	1960	m	con	secsup	dip	indet	6280	+	40
9	1962	m	con	primaria	dip	indet	13080	=	49
10	1965	f	con	primaria	autsenzadip	...	4800	+	25
11	1959	f	lib	secinf	nonlav	...	0	-	...
12	1972	f	lib	secsup	nonlav	...	0	+	...
13	...	...	...	...	...	...	...	...	...
25402	1962	m	con	secinf	autcondip	...	9864	+	60
25403	1964	f	lib	laurea	dip	indet	23352	+	32
25404	1976	m	sep	secinf	dip	indet	0	-	...
25405	1980	m	con	primaria	dip	indet	20640	++	40
25406	1982	f	con	laurea	dip	indet	10502	+	25
25407	1987	f	lib	secsup	nonlav	...	0	+	...

Esempio: effetto del fumo sul peso dei neonati (`babies`)

Per 32 neonati si sono rilevati

- il peso alla nascita (in grammi),
- la durata della gravidanza (in settimane),
- la condizione rispetto al fumo della madre (S/N).

Interessa valutare se esista una relazione tra il peso alla nascita dei neonati e la durata della gravidanza e se questa relazione cambi rispetto alla condizione madre fumatrice/non fumatrice.

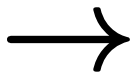
Peso	Durata gravidanza	Fumo
2940	38	S
2420	36	S
2760	39	S
2440	35	S
3301	42	S
2715	36	S
3130	39	S
2928	39	S
3446	42	S
2957	39	S
2580	38	S
3500	42	S
3200	41	S
3346	42	S
3175	41	S
2740	38	S
3130	38	N
2450	34	N
3226	40	N
2729	37	N
3410	40	N
3095	39	N
3244	39	N
2520	35	N
3523	41	N
2920	38	N
3530	42	N
3040	37	N
3322	39	N
3459	40	N
2619	35	N
2841	36	N

Gli stessi dati...

È molto spesso comodo codificare la condizione di fumatrice della madre con un numero (tipo: 1 fumatrice, 0 non fumatrice), anzichè con una lettera (S/N).

Ovviamente, i codici 0 ed 1 usati per le due condizioni sono ancora da considerarsi come etichette dei due gruppi.

Peso	Durata gravidanza	Fumo
2940	38	S
2420	36	S
2760	39	S
2440	35	S
3301	42	S
2715	36	S
3130	39	S
2928	39	S
3446	42	S
2957	39	S
2580	38	S
3500	42	S
3200	41	S
3346	42	S
3175	41	S
2740	38	S
3130	38	N
2450	34	N
3226	40	N
2729	37	N
3410	40	N
3095	39	N
3244	39	N
2520	35	N
3523	41	N
2920	38	N
3530	42	N
3040	37	N
3322	39	N
3459	40	N
2619	35	N
2841	36	N



Peso	Durata gravidanza	Fumo
2940	38	1
2420	36	1
2760	39	1
2440	35	1
3301	42	1
2715	36	1
3130	39	1
2928	39	1
3446	42	1
2957	39	1
2580	38	1
3500	42	1
3200	41	1
3346	42	1
3175	41	1
2740	38	1
3130	38	0
2450	34	0
3226	40	0
2729	37	0
3410	40	0
3095	39	0
3244	39	0
2520	35	0
3523	41	0
2920	38	0
3530	42	0
3040	37	0
3322	39	0
3459	40	0
2619	35	0
2841	36	0

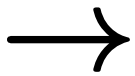
Livelli di fosfato nel plasma (`cholesterol`)

Misurazioni del livello di fosfato inorganico (mg/dl) nel plasma di soggetti obesi iperglicemici (OI), obesi non iperglicemici (ON) e di controllo (C) a un'ora dalla somministrazione di un test standard per l'assorbimento del glucosio.

OI	ON	C
2.3	3.0	3.0
4.1	4.1	2.6
4.2	3.9	3.1
4.0	3.1	2.2
4.6	3.3	2.1
4.6	2.9	2.4
3.8	3.3	2.8
5.2	3.9	3.4
3.1		2.9
3.7		2.6
3.8		3.1
		3.2

OI	ON	C
2.3	3.0	3.0
4.1	4.1	2.6
4.2	3.9	3.1
4.0	3.1	2.2
4.6	3.3	2.1
4.6	2.9	2.4
3.8	3.3	2.8
5.2	3.9	3.4
3.1		2.9
3.7		2.6
3.8		3.1
		3.2

Gli stessi dati possono essere rappresentati in forma di matrice.

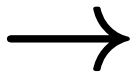


Invece di avere una variabile e tre gruppi di osservazioni avremo due variabili di cui la seconda rappresenta il gruppo.

Fosfato	Tipo di paziente
2.3	OI
4.1	OI
4.2	OI
4.0	OI
4.6	OI
4.6	OI
3.8	OI
5.2	OI
3.1	OI
3.7	OI
3.8	OI
3.0	ON
4.1	ON
3.9	ON
3.1	ON
3.3	ON
2.9	ON
3.3	ON
3.9	ON
3.0	C
2.6	C
3.1	C
2.2	C
2.1	C
2.4	C
2.8	C
3.4	C
2.9	C
2.6	C
3.1	C
3.2	C

Fosfato	Tipo di paziente
2.3	OI
4.1	OI
4.2	OI
4.0	OI
4.6	OI
4.6	OI
3.8	OI
5.2	OI
3.1	OI
3.7	OI
3.8	OI
3.0	ON
4.1	ON
3.9	ON
3.1	ON
3.3	ON
2.9	ON
3.3	ON
3.9	ON
3.0	C
2.6	C
3.1	C
2.2	C
2.1	C
2.4	C
2.8	C
3.4	C
2.9	C
2.6	C
3.1	C
3.2	C

Anche qui possiamo provare a codificare il tipo di paziente tramite 0 e 1 (come per il fumo), ma ora abbiamo 3 possibili modalità: obesi iperglicemici (OI), obesi non iperglicemici (ON) e di controllo (C).



Possiamo immaginare di sostituire la variabile Tipo di paziente con 3 variabili: Tipo OI (si/no), Tipo ON (si/no), Tipo C (si/no).

Fosfato	Tipo OI	Tipo ON	Tipo C
2.3	1	0	0
4.1	1	0	0
4.2	1	0	0
4.0	1	0	0
4.6	1	0	0
4.6	1	0	0
3.8	1	0	0
5.2	1	0	0
3.1	1	0	0
3.7	1	0	0
3.8	1	0	0
3.0	0	1	0
4.1	0	1	0
3.9	0	1	0
3.1	0	1	0
3.3	0	1	0
2.9	0	1	0
3.3	0	1	0
3.9	0	1	0
3.0	0	0	1
2.6	0	0	1
3.1	0	0	1
2.2	0	0	1
2.1	0	0	1
2.4	0	0	1
2.8	0	0	1
3.4	0	0	1
2.9	0	0	1
2.6	0	0	1
3.1	0	0	1
3.2	0	0	1

- A ben vedere, la nostra codifica è ridondante, perché sappiamo che, nel nostro studio, se un soggetto non è né obeso iperglicemico (OI) né obeso non iperglicemico (ON), non può essere altro che un soggetto di controllo (C).
- Allora ci bastano 2 sole delle 3 variabili per codificare il Tipo di paziente, per esempio Tipo OI (si/no) e Tipo ON (si/no). Sappiamo che se valgono entrambe no, il paziente è un controllo.

- A ben vedere, la nostra codifica è ridondante, perché sappiamo che, nel nostro studio, se un soggetto non è né obeso iperglicemico (OI) né obeso non iperglicemico (ON), non può essere altro che un soggetto di controllo (C).
- Allora ci bastano 2 sole delle 3 variabili per codificare il Tipo di paziente, per esempio Tipo OI (si/no) e Tipo ON (si/no). Sappiamo che se valgono entrambe no, il paziente è un controllo.
- Poi associamo alla modalità si il numero 1.

Fosfato	Tipo OI	Tipo ON
2.3	1	0
4.1	1	0
4.2	1	0
4.0	1	0
4.6	1	0
4.6	1	0
3.8	1	0
5.2	1	0
3.1	1	0
3.7	1	0
3.8	1	0
3.0	0	1
4.1	0	1
3.9	0	1
3.1	0	1
3.3	0	1
2.9	0	1
3.3	0	1
3.9	0	1
3.0	0	0
2.6	0	0
3.1	0	0
2.2	0	0
2.1	0	0
2.4	0	0
2.8	0	0
3.4	0	0
2.9	0	0
2.6	0	0
3.1	0	0
3.2	0	0

Variabili indicatrici

- Le variabili 0/1 che abbiamo creato per codificare con semplicità variabili categoriali vengono chiamate variabili **indicatrici** (perché indicano la presenza/assenza di una modalità) o variabili **dummy** (perché non sono delle vere e proprie variabili).
- Per codificare una variabile categoriale con k modalità abbiamo bisogno di $k - 1$ variabili indicatrici.
- La scelta di usare 0 e 1 è ovviamente convenzionale.

Titanic

LA DOMENICA DEL CORRIERE

Supplemento settimanale illustrato del nuovo CORRIERE DELLA SERA - Spedizione in abbonamento postale - Gruppo 2°

Anno 61 — N. 7 15 Febbraio 1959 L. 40.—

Come il "Titanic". La motonave danese "Hans Hedtoft" di 2875 tonnellate, mentre si trovava a circa 37 miglia a sud di Capo Farvel (Groenlandia), è stata investita da un gigantesco iceberg. Di ritorno dal suo viaggio inaugurale, essa faceva rotta su Copenhagen con 55 passeggeri e 40 uomini di equipaggio. L'ultimo messaggio diceva: "Affondiamo lentamente. Chiediamo assistenza immediata". Navi da costa e da pesca si sono dirette verso la "Hedtoft", numerosi aerei decollati da basi della Groenlandia hanno perlustrato la zona da cui era stato segnalato il sinistro. Ma invano. Della motonave nessuna traccia. A distanza di 47 anni si è rinnovata la tragedia del "Titanic". (Disegno di Walter Molino)

Il transatlantico britannico *RMS Titanic* affonda a seguito della collisione con un *iceberg* nella notte tra il 14 e il 15 aprile 1912.

Delle 2201 persone a bordo tra passeggeri ed equipaggio, sopravvivono solo 711.

Tra le polemiche che seguono al naufragio c'è chi sostiene che i passeggeri di III classe vennero trascurati nelle operazioni di evacuazioni, dando preferenza ai "ricchi".

Titanic

	Deceduto	Sopravv.
I Cl.	122	203
II Cl.	167	118
III Cl.	528	178
Equipaggio	673	212

I dati a disposizione sono riassumibili in una tabella a doppia entrata, in cui si riporta il numero di sopravvissuti e di deceduti a seconda della classe di appartenenza.

Il sospetto per cui i passeggeri di III classe vennero trascurati si traduce nel dire che le due caratteristiche osservate: sopravvivenza e classe, sono legate.

Il disastro del Titanic

Nome	Passeggero (tipologia)	Sopravvivenza
nome 1	II	sopravvissuto
nome 2	III	non sopravvissuto
nome 3	I	non sopravvissuto
⋮	⋮	⋮
nome 2201	equipaggio	sopravvissuto

Indice

Distribuzione statistica disaggregata

Si consideri un collettivo statistico di N unità, dove si sia osservato il carattere Y. Si chiama **distribuzione statistica disaggregata** secondo il carattere Y l'insieme delle osservazioni (rappresentate da numeri o da espressioni verbali) relative alle N unità del collettivo. In simboli, la distribuzione disaggregata sarà indicata come

$$y_1, y_2, \ldots, y_N$$

dove y_1 è l'osservazione relativa all'unità identificata dal numero 1, y_2 l'osservazione relativa all'unità identificata dal numero 2 e così via.

Non consente una facile visione d'insieme.

Distribuzione statistica disaggregata (continua)

Ad esempio per il carattere *Genere la distribuzione statistica disaggregata*

Maschio Maschio Maschio Maschio Maschio Femmina Femmina Femmina Maschio Maschio Femmina Femmina Maschio Maschio Femmina Femmina Femmina Maschio Maschio Femmina Femmina Maschio Femmina Femmina Maschio Femmina Femmina Femmina Femmina Maschio Maschio Maschio Maschio Femmina Maschio Maschio Femmina Maschio Maschio Femmina Femmina Maschio Femmina Maschio Femmina Maschio Maschio Femmina Maschio Maschio Maschio Maschio Maschio Maschio Maschio Femmina Maschio Maschio Femmina Maschio Femmina Maschio Femmina Femmina Femmina Maschio Maschio Femmina Femmina Maschio Maschio Femmina Maschio Maschio Maschio Maschio Femmina Maschio Femmina Maschio Femmina Maschio Maschio Femmina Maschio Femmina Femmina Maschio Femmina Femmina Maschio Femmina Maschio Femmina

Distribuzione di frequenza assoluta

Si consideri ancora Y. Si chiama **distribuzione di frequenza assoluta** la lista delle modalità osservate accompagnata dal numero di volte in cui queste vengono osservate, ossia accompagnata dalle rispettive **frequenze assolute**.

È molto facile ottenere distribuzioni di frequenza assoluta per caratteri qualitativi e quantitativi discreti. In presenza di caratteri quantitativi continui (o anche discreti, se assumono tantissime modalità), abbiamo bisogno di qualche operazione preliminare.

Esempio: distribuzione di frequenza del *genere*

Modalità	Frequenza assoluta
Femmina	177
Maschio	199

Esempio: distribuzione di frequenza del *genere letterario preferito*

Modalità	Frequenza assoluta
Altro	81
Comico/umoristico	6
Fantascienza	16
Fantasy	32
Giallo/noir/thriller	73
Orrore	11
Psicologico	58
Romantico	62
Storico	33

Esempio: distribuzione di frequenza delle *ore di sonno*

Modalità	Frequenza assoluta
5	7
6	32
7	147
8	165
9	19
10	2
12	1

Esempio: distribuzione di frequenza dell'*Altezza*

Per l'altezza conviene definire delle classi.

Modalità	Frequenza assoluta
147	1
155	1
157	2
158	11
159	2
160	19
161	6
162	10
163	7
164	8
165	25
166	3
167	7
168	14
169	7
170	30
171	2
172	11
173	12
174	6
175	26
176	7
177	6
178	17
179	7
180	27

Esempio: distribuzione di frequenza dell'*Altezza*

Per l'altezza conviene definire delle classi.

Modalità	Frequenza assoluta
147	1
155	1
157	2
158	11
159	2
160	19
161	6
162	10
163	7
164	8
165	25
166	3
167	7
168	14
169	7
170	30
171	2
172	11
173	12
174	6
175	26
176	7
177	6
178	17
179	7
180	27

Modalità	Frequenza assoluta
[145,160]	36
(160,165]	56
(165,170]	61
(170,175]	57
(175,180]	64
(180,185]	51
(185,190]	39
(190,195]	9

Esempio: distribuzione di frequenza dell'*Altezza*

Per l'altezza conviene definire delle classi.

Modalità	Frequenza assoluta
147	1
155	1
157	2
158	11
159	2
160	19
161	6
162	10
163	7
164	8
165	25
166	3
167	7
168	14
169	7
170	30
171	2
172	11
173	12
174	6
175	26
176	7
177	6
178	17
179	7
180	27

Modalità	Frequenza assoluta
[145,160]	36
(160,165]	56
(165,170]	61
(170,175]	57
(175,180]	64
(180,185]	51
(185,190]	39
(190,195]	9

Modalità	Frequenza assoluta
[145,185]	325
(185,195]	48

Classi di differenti lunghezze

Può capitare, o per scelta (si vuole fornire informazioni più dettagliate su parte della distribuzione), o per necessità (i dati sono già stati raggruppati in classi da qualcuno), di costruire delle classi utilizzando intervalli di lunghezza differente.

In questo caso è conveniente definire anche la **densità** di frequenza.

La densità è definita come:

$$\left(\begin{array}{c}\text{densità}\\ \text{di una classe}\end{array}\right) = \frac{\text{frequenza assoluta di } Y \text{ sull'intervallo}}{\text{lunghezza dell'intervallo}}.$$

Esempio: amici di Facebook

Modalità	Freq. ass.
0	1
15	1
30	1
40	1
41	1
50	1
80	3
99	1
100	1
120	1
123	1
124	1
130	1
150	1
173	1
183	1
200	4
215	1
229	1
236	1
240	1
242	1
247	1
250	4
270	1
272	1
280	1
281	1
289	2
290	1
300	20
310	1
315	1
320	3
322	1
328	2

Modalità	Freq. ass.
700	8
715	1
719	1
723	1
724	1
725	1
750	1
755	1
760	2
771	1
774	1
790	1
796	1
800	8
823	1
826	1
850	2
859	1
860	1
864	1
865	1
888	1
900	4
911	1
913	1
925	1
927	1
952	1
970	1
976	1
987	1
1000	14
1003	1
1017	1
1018	1
1032	1

Esempio: amici di Facebook

Modalità	Freq. ass.
[0,500]	130
(500,1000]	116
(1000,1500]	48
(1500,2000]	14
(2000,2500]	10
(2500,3000]	5
(3000,100000]	3

Modalità	Freq. ass.
0	1
15	1
30	1
40	1
41	1
50	1
80	3
99	1
100	1
120	1
123	1
124	1
130	1
150	1
173	1
183	1
200	4
215	1
229	1
236	1
240	1
242	1
247	1
250	4
270	1
272	1
280	1
281	1
289	2
290	1
300	20
310	1
315	1
320	3
322	1
328	2

Modalità	Freq. ass.
700	8
715	1
719	1
723	1
724	1
725	1
750	1
755	1
760	2
771	1
774	1
790	1
796	1
800	8
823	1
826	1
850	2
859	1
860	1
864	1
865	1
888	1
900	4
911	1
913	1
925	1
927	1
952	1
970	1
976	1
987	1
1000	14
1003	1
1017	1
1018	1
1032	1

Densità di frequenza

Modalità	Freq. ass.	Ampiezza classe	Densità
[0,100]	11	100	11/100=0.11
(100,200]	11	100	11/100=0.11
(200,300]	37	100	37/100=0.37
(300,400]	39	100	39/100=0.39
(400,500]	32	100	32/100=0.32
(500,1000]	116	500	116/500=0.232
(1000,10000]	79	9000	79/9000=0.00878

La densità ci dice il numero atteso di unità statistiche per ogni unità di misura della variabile. Nella prima classe, per esempio, ci aspettiamo di osservare 11 persone in un intervallo di 100 unità, nella penultima classe ci aspettiamo di vedere 116 unità ogni 500 .

Esempio: ore di sonno, maschi e femmine

Alle volte è utile considerare la distribuzione di una variabile – qui le ore di sonno – per sottogruppi corrispondenti ai valori di un'altra variabile – qui il genere.

Femmine

Modalità	Freq. ass.
5	3
6	14
7	71
8	82
9	6

Maschi

Modalità	Freq. ass.
5	4
6	18
7	76
8	83
9	13
10	2
12	1

Si parla in questo caso di distribuzioni di frequenze condizionate.

Esempio: ore di sonno, maschi e femmine

Alle volte è utile considerare la distribuzione di una variabile – qui le ore di sonno – per sottogruppi corrispondenti ai valori di un'altra variabile – qui il genere.

Femmine

Modalità	Freq. ass.
5	3
6	14
7	71
8	82
9	6

Maschi

Modalità	Freq. ass.
5	4
6	18
7	76
8	83
9	13
10	2
12	1

Totale

Modalità	Freq. ass.
5	7
6	32
7	147
8	165
9	19
10	2
12	1

Si parla in questo caso di distribuzioni di frequenze condizionate.

Esempio: ore di sonno, maschi e femmine

Alle volte è utile considerare la distribuzione di una variabile – qui le ore di sonno – per sottogruppi corrispondenti ai valori di un'altra variabile – qui il genere.

Modalità	Freq. ass.		
	Femmine	Maschi	Totale
5	3	4	7
6	14	18	32
7	71	76	147
8	82	83	165
9	6	13	19
10	0	2	2
12	0	1	1

Si parla in questo caso di distribuzioni di frequenze condizionate.

Distribuzioni condizionate

- Se indichiamo in modo generico con
  - Y il carattere che stiamo studiando (le ore di sonno, per esempio)
  - X il carattere tramite cui estraiamo le unità statistiche da considerare nell'analisi (il genere, nel nostro caso)

  si dice variabile Y condizionata a $X = x$ e si indica $Y|X = x$ la restrizione di Y al sottoinsieme $X = x$.
- La distribuzione della variabile $Y|X = x$ viene normalmente detta la **distribuzione di Y condizionata a $X = x$** o, equivalentemente, la **distribuzione di Y dato $X = x$**.
- Si osservi che esiste una distribuzione condizionata (di Y dato X) per ogni modalità di X.
- La distribuzione della variabile Y senza distinzione per condizione rispetto a X è detta **distribuzione marginale**.

Esempio: altezza, maschi e femmine

Nel caso di variabili quantitative continue, che come si è detto si possono dividere in intervalli, le cose funzionano in modo analogo.

Femmine

Modalità	Freq. ass.
[145,160]	36
(160,165]	56
(165,170]	50
(170,175]	29
(175,180]	4
(180,185]	0
(185,190]	0
(190,195]	0

Maschi

Modalità	Freq. ass.
[145,160]	0
(160,165]	0
(165,170]	11
(170,175]	28
(175,180]	60
(180,185]	51
(185,190]	39
(190,195]	9

Frequenze relative

Dividendo una frequenza assoluta per il numero totale di unità statistiche (N nel nostro caso) otteniamo le cosiddette **frequenze relative**, ovvero

$$\left(\begin{array}{c}\text{frequenze}\\ \text{relative}\end{array}\right) = \frac{\left(\begin{array}{c}\text{frequenze}\\ \text{assolute}\end{array}\right)}{\left(\begin{array}{c}\text{numero totale di}\\ \text{osservazioni}\end{array}\right)}$$

Hanno il vantaggio, rispetto alle frequenze assolute, di permettere di confrontare distribuzioni di frequenza basate su numeri differenti di unità statistiche.

Frequenze relative: ore di sonno

Mod	Freq. ass.		
	F	M	Tot
5	3	4	7
6	14	18	32
7	71	76	147
8	82	83	165
9	6	13	19
10	0	2	2
12	0	1	1

Mod	Freq. rel.		
	F	M	Tot
5	0.02	0.02	0.02
6	0.08	0.09	0.09
7	0.40	0.39	0.39
8	0.47	0.42	0.44
9	0.03	0.07	0.05
10	0.00	0.01	0.01
12	0.00	0.01	0.00

Esempio: amici su Facebook

Modalità	Freq. rel.	Freq. rel.
[0,100]	11	11/326=0.034
(100,200]	11	11/326=0.034
(200,300]	37	37/326=0.113
(300,400]	39	39/326=0.12
(400,500]	32	32/326=0.098
(500,750]	66	66/326=0.202
(750,1000]	50	50/326=0.153
(1000,2000]	62	62/326=0.19
(2000,5000]	16	16/326=0.049
(5000,100000]	2	2/326=0.006

Distribuzioni di frequenza: notazione

- y_i modalità i/ classe $(c_{i-1}, c_i]$ del carattere y, $i = 1, 2, \ldots, k$ (k modalità/classi);
- n_i frequenza assoluta numero di unità statistiche che possiedono la modalità/classe y_i;
- N numero totale di osservazioni ($N = n_1 + n_2 + \cdots + n_k$);
- f_i frequenza relativa ($f_i = n_i/N$).

modalità/classe	frequenze assolute	frequenze relative
y_1	n_1	$f_1 = n_1/N$
y_2	n_2	$f_2 = n_2/N$
$\vdots$	$\vdots$	$\vdots$
y_k	n_k	$f_k = n_k/N$
Totale	N	1

Avviso generale

Ogni libro usa una propria notazione, magari diversa da quella appena introdotta.

Una altra notazione comune è, per esempio, la seguente

- y_i modalità/classe i del carattere y, $i = 1, 2, \ldots, k$ (k modalità/classi)
- f_i frequenza assoluta numero di unità statistiche che possiedono la modalità/classe y_i
- n numero totale di osservazioni ($n = f_1 + f_2 + \cdots + f_k$)
- p_i frequenza relativa ($p_i = f_i/n$)

Qualunque scelta va bene: basta definire cosa si intende con ciascun simbolo ed essere coerenti.

Il simbolo $\sum$ (sommatoria)

Cosa intendiamo per

$$N = \sum_{i=1}^{k} n_i$$

ovvero per 'Somma per i che va da 1 a k' ?

$$N = n_1 + n_2 + \cdots + n_k$$

Alcune proprietà

1. $\sum_{i=1}^{k}(y_i + x_i) = \sum_{i=1}^{k} y_i + \sum_{i=1}^{k} x_i$
2. $\sum_{i=1}^{k} a y_i = a \sum_{i=1}^{k} y_i$
3. Fate attenzione: $\sum_{i=1}^{k} a = ak$

Esercizio: $\sum_{i=1}^{k} f_i = ?$

Esercizio: esiti ammissione a Berkeley, 1973

I dati a destra rappresentano gli esiti dell'ammissione all'Università di California, Berkeley (USA) nel 1973. È riportato l'esito dell'ammissione (`Admit`), il sesso dei candidati (`Gender`) e il Dipartimento erogante il corso di studi scelto dai candidati (`Dept`).

—

È una matrice dei dati?

Quante sono le variabili rilevate?

Di che tipo sono?

Quale è il loro supporto?

Quante sono le unità statistiche?

I dati si riferiscono ad una indagine censuaria o campionaria?

Admit	Gender	Dept	Frequenza assoluta
Admitted	Male	A	512
Rejected	Male	A	313
Admitted	Female	A	89
Rejected	Female	A	19
Admitted	Male	B	353
Rejected	Male	B	207
Admitted	Female	B	17
Rejected	Female	B	8
Admitted	Male	C	120
Rejected	Male	C	205
Admitted	Female	C	202
Rejected	Female	C	391
Admitted	Male	D	138
Rejected	Male	D	279
Admitted	Female	D	131
Rejected	Female	D	244
Admitted	Male	E	53
Rejected	Male	E	138
Admitted	Female	E	94
Rejected	Female	E	299
Admitted	Male	F	22
Rejected	Male	F	351
Admitted	Female	F	24
Rejected	Female	F	317

Frequenze cumulate

- Ha senso se il carattere è ordinato: $y_1 < y_2 < \ldots < y_k$
- La frequenza assoluta (relativa) cumulata per la modalità/classe y_j è la somma delle frequenze assolute (relative) per le modalità/classi $\leq y_j$

$$F_j = f_1 + \ldots + f_j = \sum_{h=1}^{j} f_h$$

modalità/classe	frequenze cumulate assolute	frequenze cumulate relative
y_1	n_1	$F_1 = f_1$
y_2	$n_1 + n_2$	$F_2 = f_1 + f_2$
$\vdots$	$\vdots$	$\vdots$
y_j	$n_1 + \ldots + n_j$	$F_j = f_1 + \ldots + f_j$
$\vdots$	$\vdots$	$\vdots$
y_k	N	?

Esercizio: ore di sonno per notte

Si costruisca la distribuzione di frequenze cumulate per le ore di sonno per notte Partendo dalla distribuzione di frequenze assolute, abbiamo

Modalità	Frequenza assoluta	Frequenza assoluta cumulata	Frequenza relativa	Frequenza relativa cumulata
5	7	7	0.019	0.019
6	32	39	0.086	0.105
7	147	186	0.394	0.499
8	165	351	0.442	0.941
9	19	370	0.051	0.992
10	2	372	0.005	0.997
12	1	373	0.003	1.000

Indice

Funzione di ripartizione empirica

La distribuzione di frequenze relative cumulate è collegata ad una importante rappresentazione dell'andamento di una variabile quantitativa, ossia la **funzione di ripartizione empirica**.

$$\left(\begin{array}{c}\text{funzione di}\\ \text{ripartizione empirica}\\ \text{calcolata in } y\end{array}\right) = \frac{\left(\begin{array}{c}\text{numero di}\\ \text{osservazioni minori o}\\ \text{uguali a } y\end{array}\right)}{\left(\begin{array}{c}\text{numero totale di}\\ \text{osservazioni}\end{array}\right)}$$

Funzione di ripartizione empirica

Formalizzando, detto Y il carattere oggetto di studio, la funzione di ripartizione empirica calcolata a partire dal campione $(y_1, y_2, \ldots, y_N)$ è la funzione

$$F_Y(y) = \frac{\#\{Y \leq y\}}{N}.$$

Il dominio della funzione è dato da $\mathbb{R}$; il codominio è l'intervallo $[0, 1]$.

Esempio: ore di sonno

Costruiamo la funzione di ripartizione empirica per le ore di sonno.

Mod.	Freq. ass.	Freq. ass. cum.
5	7	7
6	32	39
7	147	186
8	165	351
9	19	370
10	2	372
12	1	373

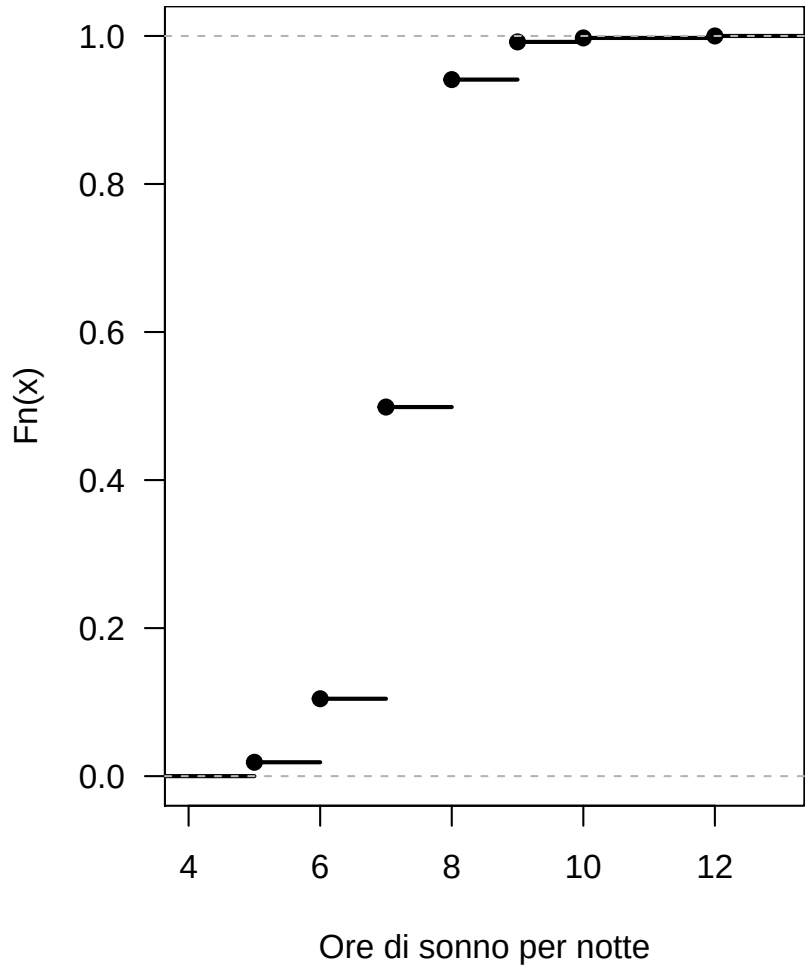
Partendo dalla distribuzione di frequenze relative cumulate, la funzione è così definita.

$$F_Y(y) = \begin{cases} 0 & y < 5 \\ 7/373 & 5 \leq y < 6 \\ 39/373 & 6 \leq y < 7 \\ 186/373 & 7 \leq y < 8 \\ 351/373 & 8 \leq y < 9 \\ 370/373 & 9 \leq y < 10 \\ 372/373 & 10 \leq y < 12 \\ 1 & y \geq 12 \end{cases}$$

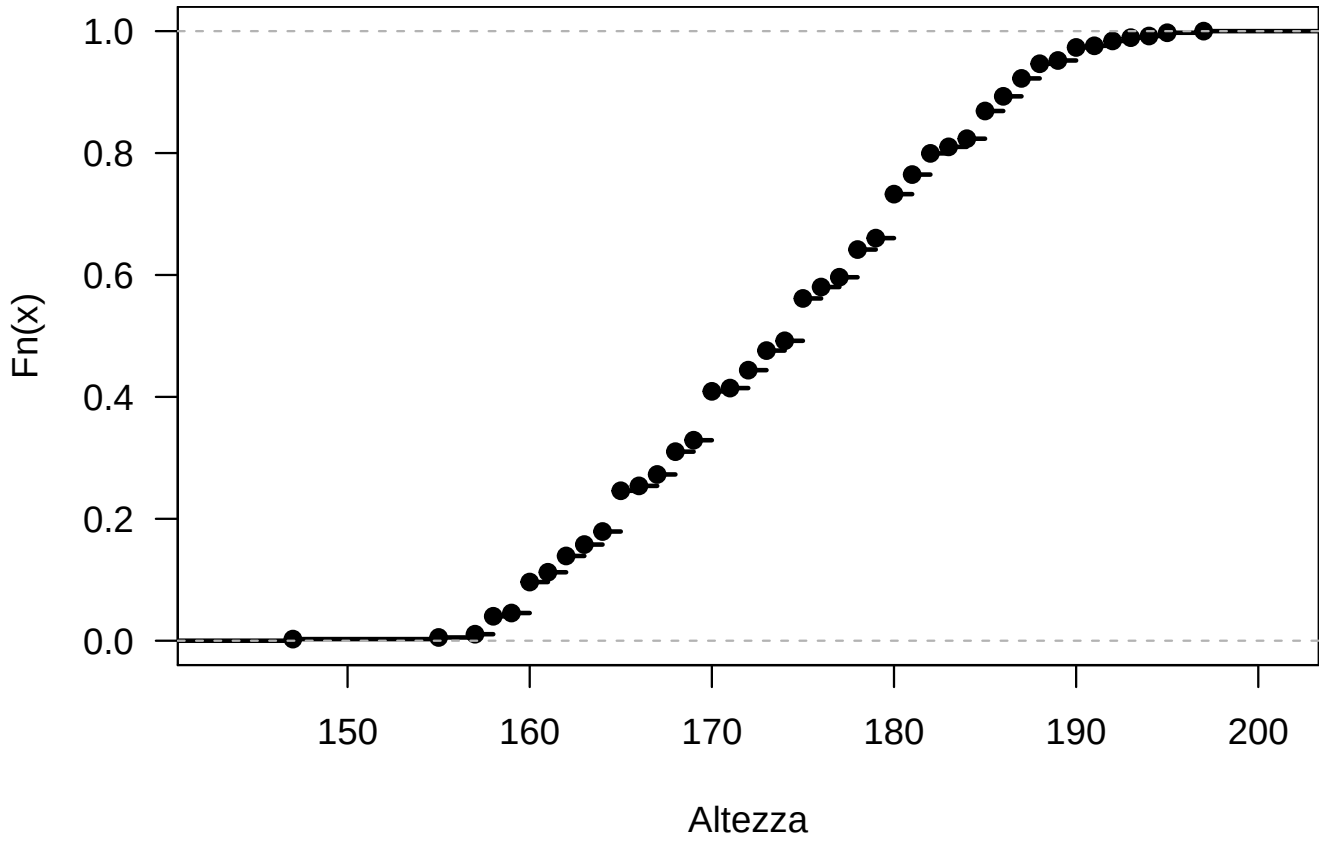
Funzione di ripartizione empirica: rappresentazione grafica

La funzione di ripartizione empirica può essere rappresentata graficamente.

$$F_Y(y) = \begin{cases} 0 & y < 5 \\ 7/373 & 5 \leq y < 6 \\ 39/373 & 6 \leq y < 7 \\ 186/373 & 7 \leq y < 8 \\ 351/373 & 8 \leq y < 9 \\ 370/373 & 9 \leq y < 10 \\ 372/373 & 10 \leq y < 12 \\ 1 & y \geq 12 \end{cases}$$



Funzione di ripartizione empirica: altezze



Indice

Finalmente un grafico!

Possiamo cercare di visualizzare le distribuzioni di frequenza, rappresentando in qualche modo ciascuna modalità del carattere con la relativa frequenza.

Esempio: il vostro rapporto con Sanremo.

Modalità	freq. ass.	freq. rel.
Non visto	252	0.67
Visto e non piaciuto	34	0.09
Visto e piaciuto	88	0.24

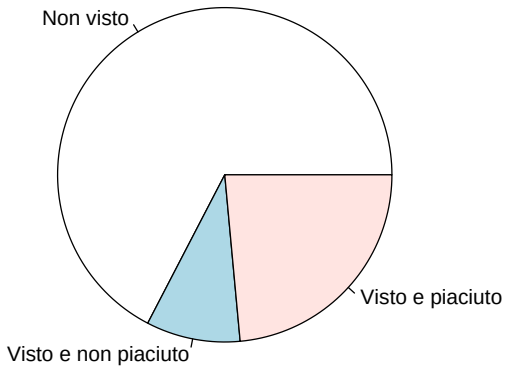


Diagramma a torta

Il grafico è costruito rappresentando ogni modalità con una fetta di torta di superficie proporzionale alla sua frequenza:

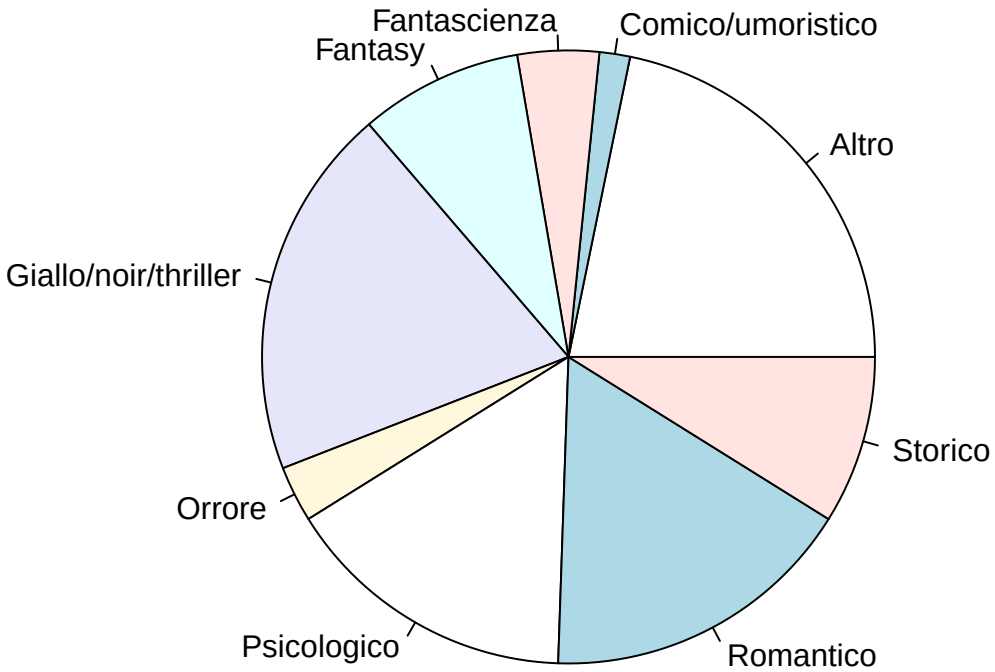
$$\textit{angolo} = 360 \cdot \textit{frequenza assoluta}/N$$

o

$$\textit{angolo} = 360 \cdot \textit{frequenza relativa}$$

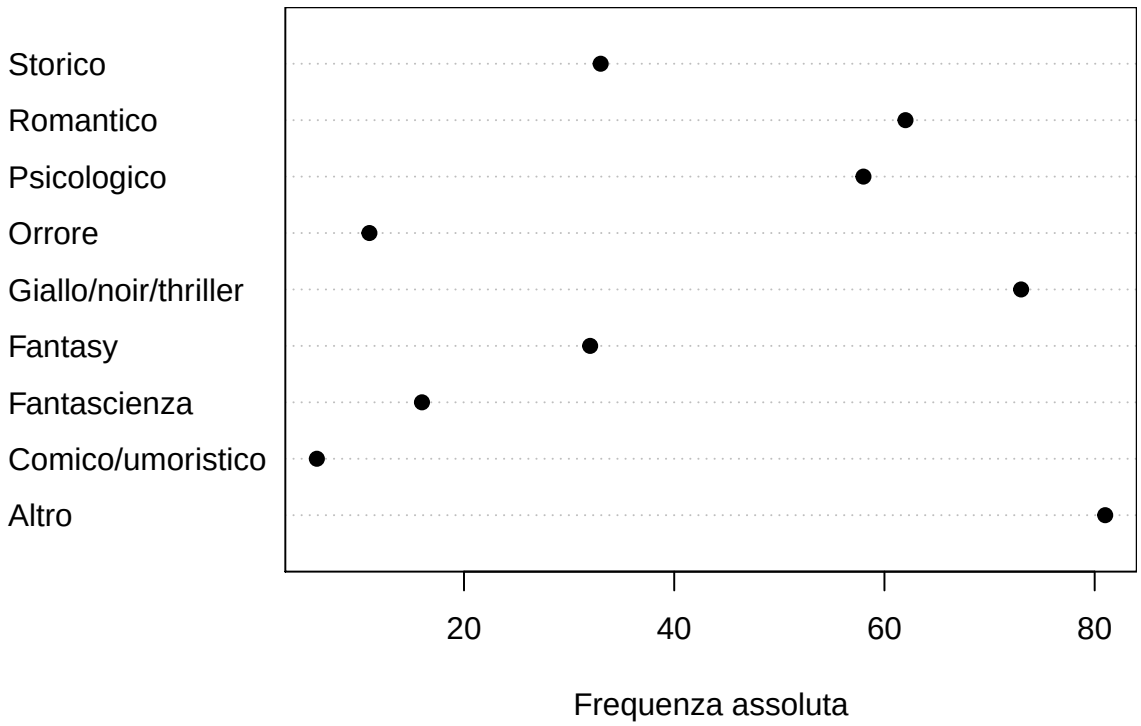
Finalmente un grafico!

Esempio: generi letterari preferiti



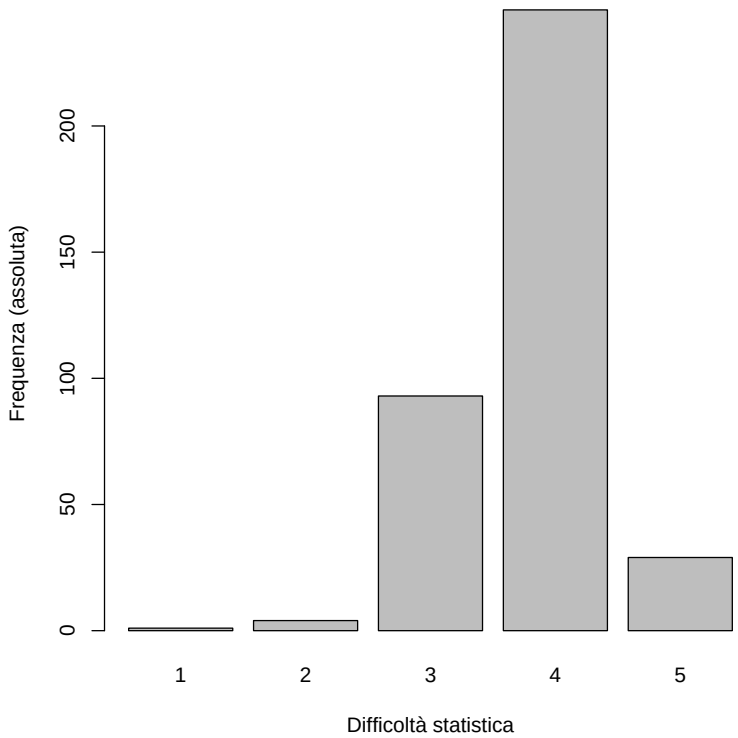
Finalmente un grafico!

Esempio: generi letterari preferiti



Notiamo che, se la variabile non è ordinale, l'ordine delle modalità nel grafico è arbitrario.

Esempio: quanto è difficile statistica

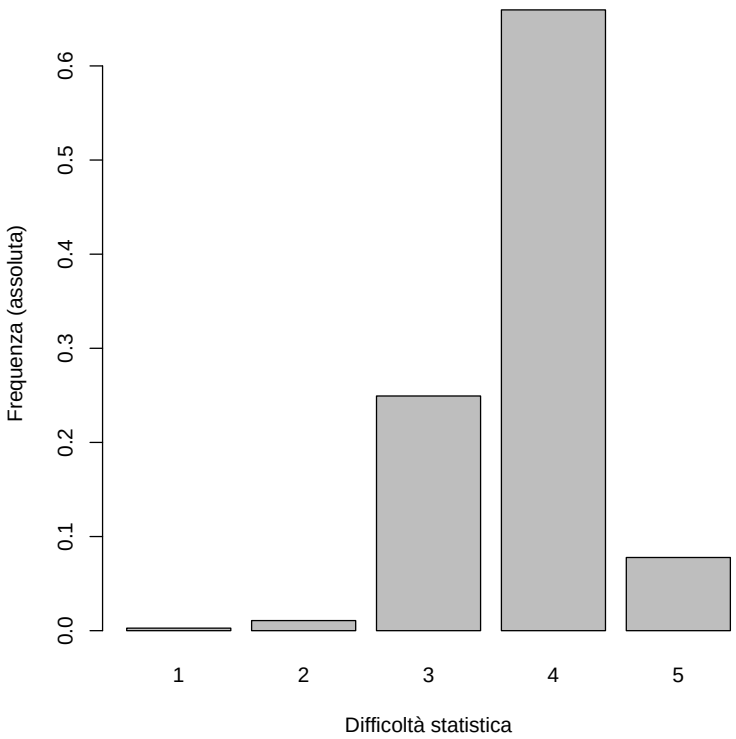


Il grafico è stato costruito ponendo

$$\text{asse}x = \begin{pmatrix}\text{modalità riportate}\\ \text{nella distribuzione}\\ \text{di frequenza}\end{pmatrix}$$

$$(\text{altezza barre}) = (\text{frequenze assolute})$$

Esempio: quanto è difficile statistica



Il grafico è stato costruito ponendo

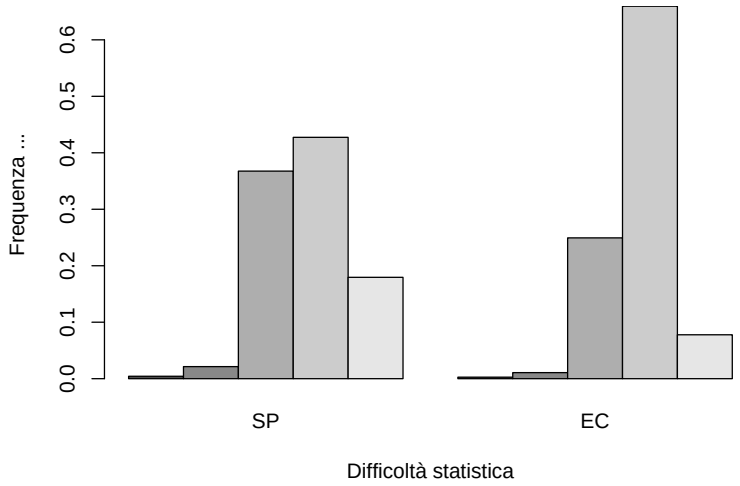
$$\text{asse}x = \begin{pmatrix}\text{modalità riportate}\\ \text{nella distribuzione}\\ \text{di frequenza}\end{pmatrix}$$

$$(\text{altezza barre}) = (\text{frequenze assolute})$$

Lo stesso grafico si può fare per le frequenze relative

$$(\text{altezza barre}) = (\text{frequenze relative})$$

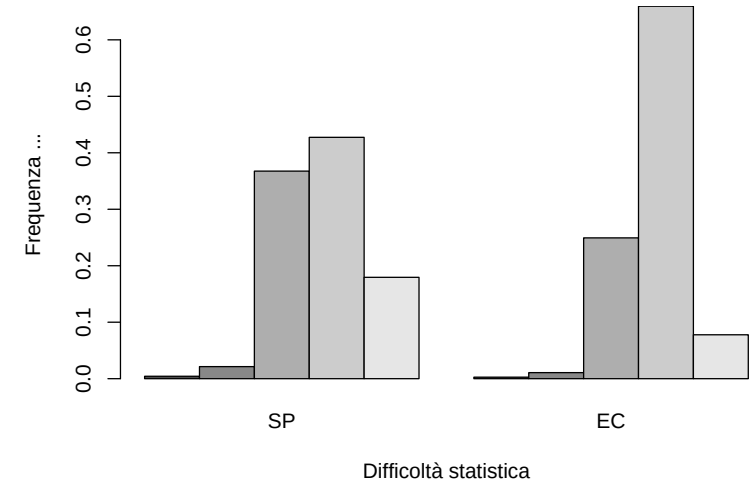
Diagrammi a barre e confronto: difficoltà e CdS

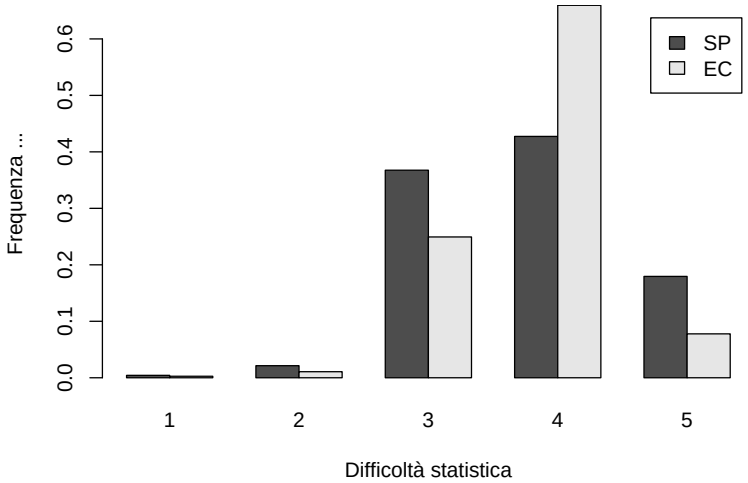


Confrontiamo le risposte date sulla difficoltà dai frequentanti Economia (EC) e dai frequentanti Scienze Politiche (SP).

Che frequenze sono?

Diagrammi a barre e confronto: difficoltà e CdS

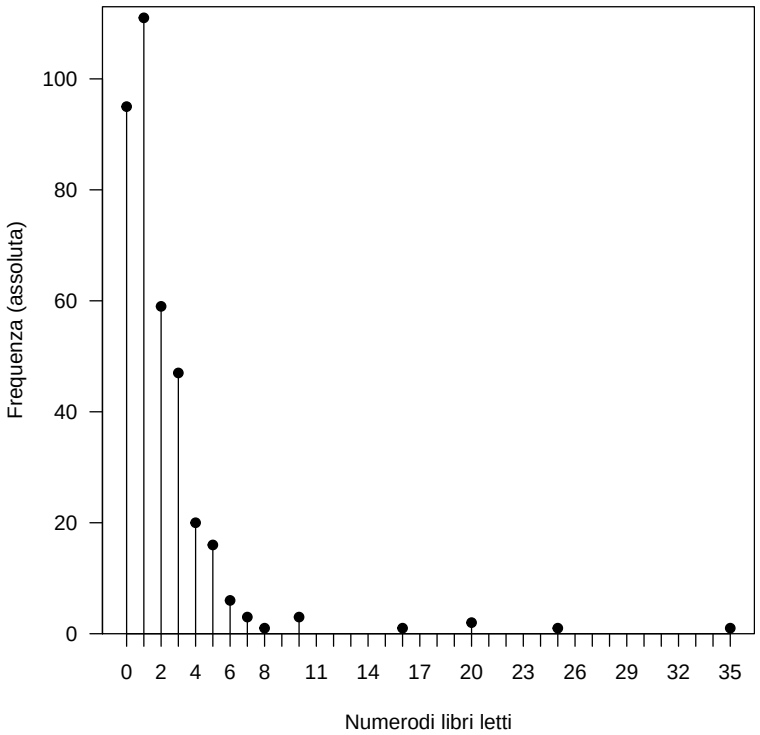




Confrontiamo le risposte date sulla difficoltà dai frequentanti Economia (EC) e dai frequentanti Scienze Politiche (SP).

Che frequenze sono?

Esempio: numero di libri letti



Il grafico è stato costruito ponendo

$$\text{asse}x = \begin{pmatrix}\text{modalità riportate}\\ \text{nella distribuzione}\\ \text{di frequenza}\end{pmatrix}$$

$$(\text{altezza barre}) = (\text{frequenze assolute})$$

Esempio: altezze

[145,160]	36
(160,165]	56
(165,170]	61
(170,175]	57
(175,180]	64
(180,185]	51
(185,190]	39
(190,195]	9

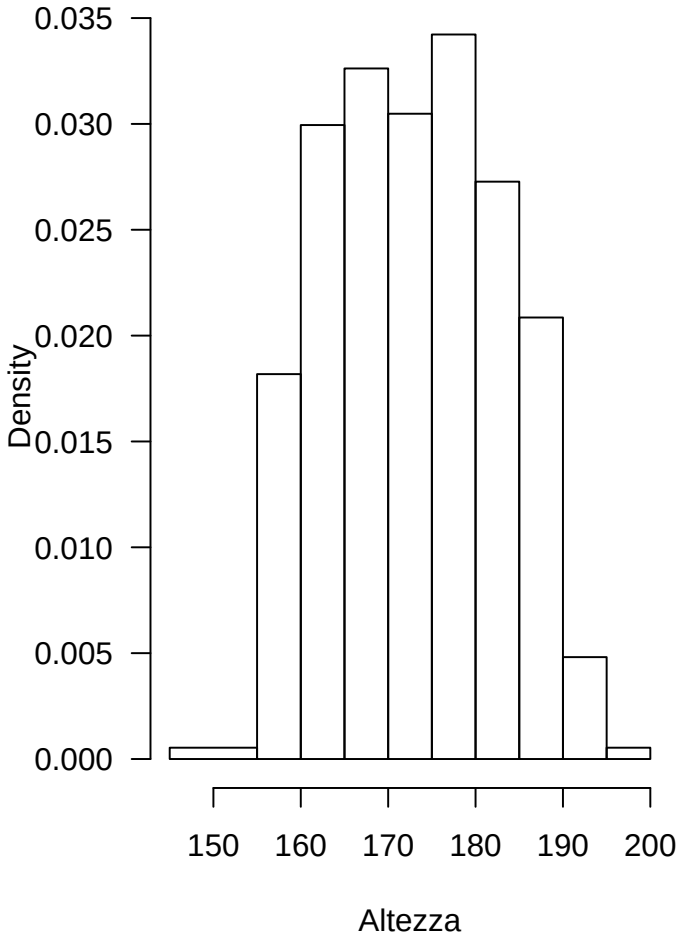
Il grafico è stato costruito ponendo

$$\text{(base rettangoli)} = \begin{pmatrix} \text{intervallini riportati} \\ \text{nella } 1^\circ \text{ colonna} \\ \text{della distribuzione} \\ \text{di frequenza} \end{pmatrix}$$

$$\text{(area rettangoli)} \propto \text{(frequenze assolute)}$$

Il simbolo $\propto$ significa "proporzionale a".

Esempio: altezze



Il grafico è stato costruito ponendo

$$(\text{base rettangoli}) = \begin{pmatrix} \text{intervallini riportati} \\ \text{nella } 1^\circ \text{ colonna} \\ \text{della distribuzione} \\ \text{di frequenza} \end{pmatrix}$$

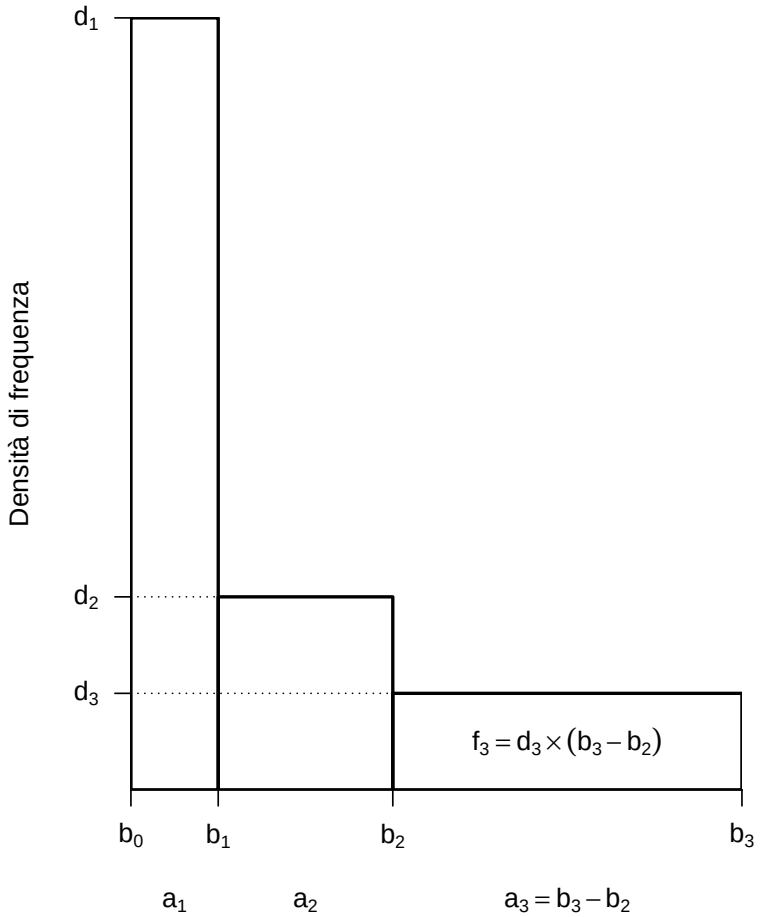
$$(\text{area rettangoli}) \propto (\text{frequenze assolute})$$

Il simbolo $\propto$ significa "proporzionale a".

Essendo l'area dei rettangoli uguale a base$\times$altezza, se le gli intervalli hanno uguale ampiezza, di fatto l'altezza coincide con (o è proporzionale a) la frequenza assoluta:

$$(\text{altezza rettangoli}) = (\text{frequenze assolute})$$

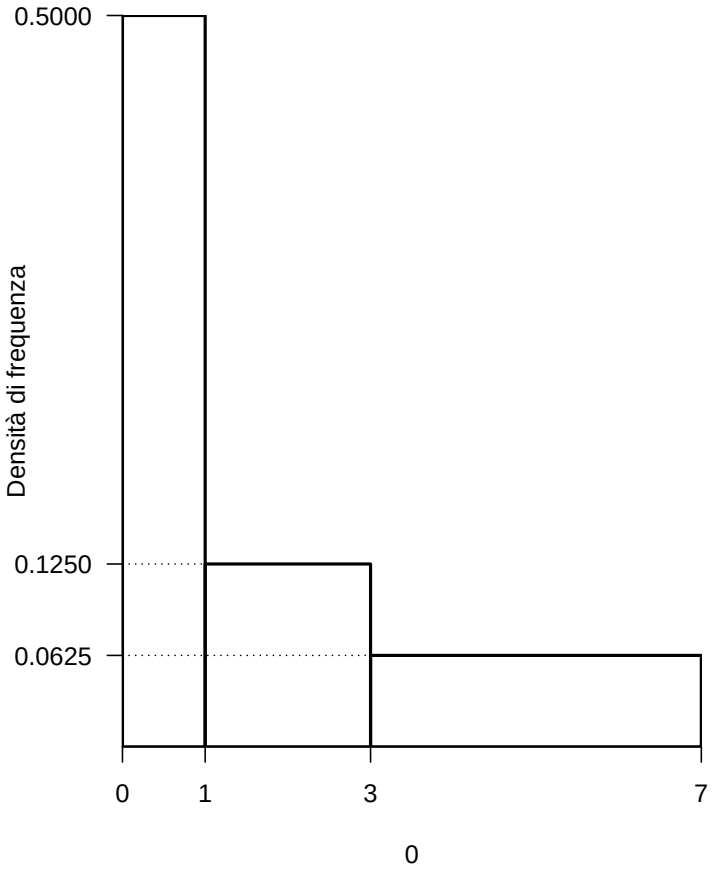
Costruzione di un istogramma



i	b_{i-1}	b_i	a_i	n_i	f_i	d_i
1	0	1	1	20	0.50	0.500
2	1	3	2	10	0.25	0.125
3	3	7	4	10	0.25	0.062

$$
\begin{aligned}
a_i &= b_i - b_{i-1} \\
f_i &= \frac{n_i}{\sum_{i=1}^{3} n_i} \\
d_i &= \frac{f_i}{a_i} = \frac{f_i}{b_i - b_{i-1}}
\end{aligned}
$$

Costruzione di un istogramma

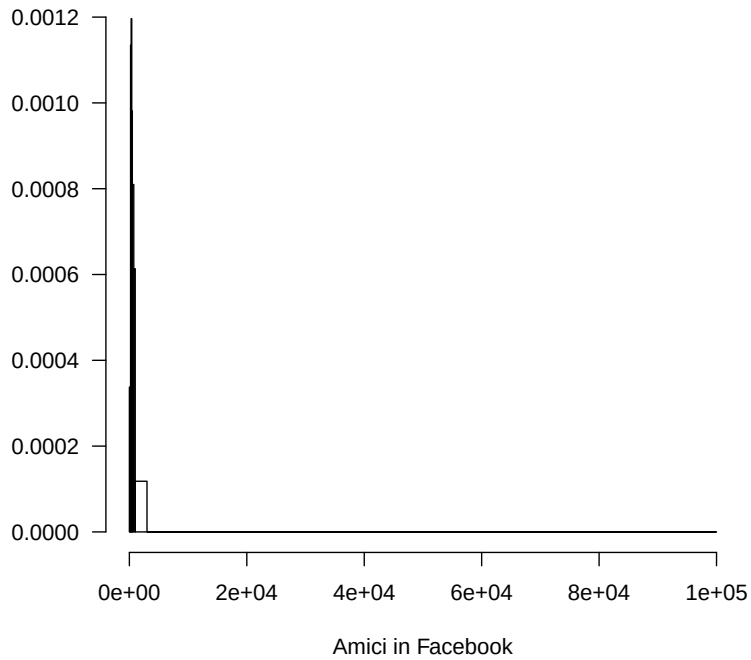


i	b_{i-1}	b_i	a_i	n_i	f_i	d_i
1	0	1	1	20	0.50	0.500
2	1	3	2	10	0.25	0.125
3	3	7	4	10	0.25	0.062

$$
\begin{aligned}
a_i &= b_i - b_{i-1} \\
f_i &= \frac{n_i}{\sum_{i=1}^{3} n_i} \\
d_i &= \frac{f_i}{a_i} = \frac{f_i}{b_i - b_{i-1}}
\end{aligned}
$$

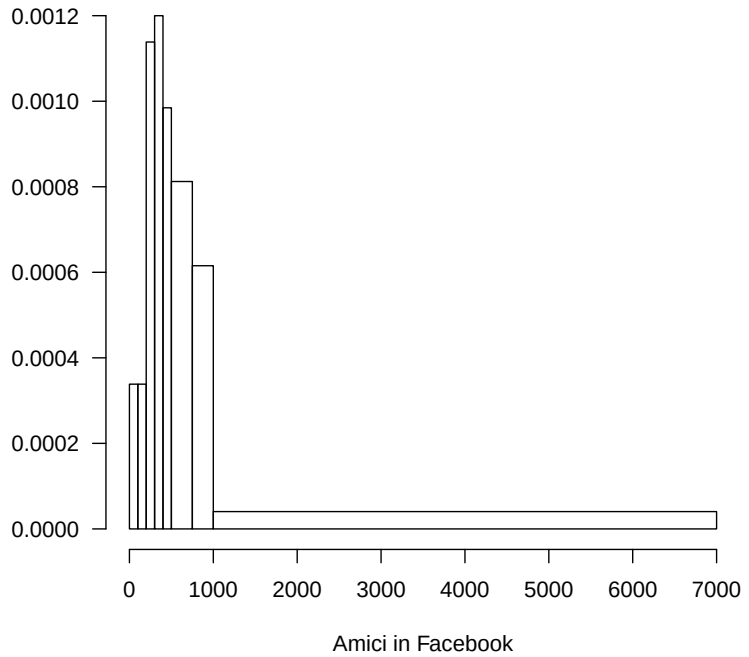
Istogramma: esempio

Modalità	Freq. rel.	Ampiezza classe	Densità
[0,100]	0.0337	100	0.000340
(100,200]	0.0337	100	0.000340
(200,300]	0.1135	100	0.001140
(300,400]	0.1196	100	0.001200
(400,500]	0.0982	100	0.000980
(500,750]	0.2024	250	0.000810
(750,1000]	0.1534	250	0.000610
(1000,2000]	0.1902	1000	0.000190
(2000,7000]	0.0522	5000	0.000010
(7000,100000]	0.0031	93000	0.000000



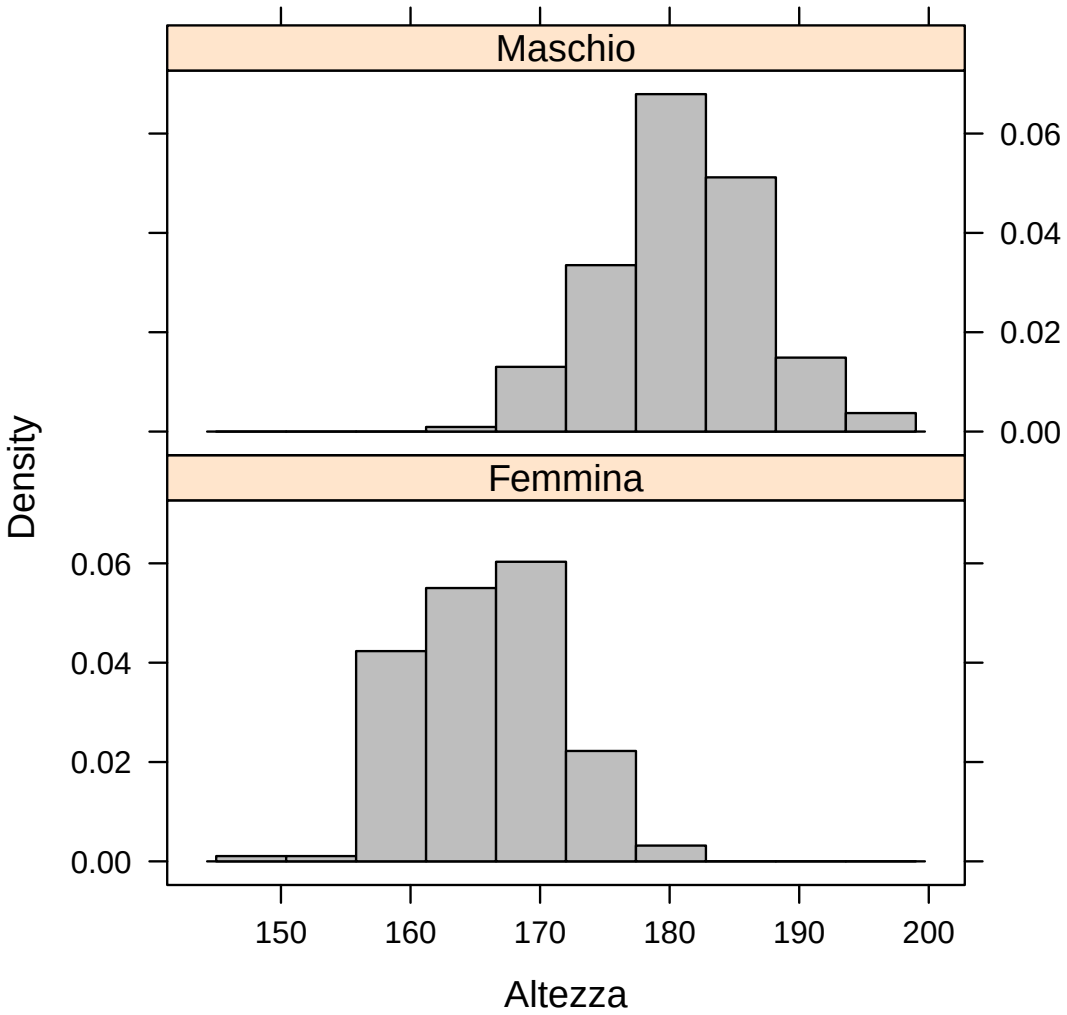
Istogramma: esempio

Modalità	Freq. rel.	Ampiezza classe	Densità
[0,100]	0.0337	100	0.000340
(100,200]	0.0337	100	0.000340
(200,300]	0.1135	100	0.001140
(300,400]	0.1196	100	0.001200
(400,500]	0.0982	100	0.000980
(500,750]	0.2024	250	0.000810
(750,1000]	0.1534	250	0.000610
(1000,2000]	0.1902	1000	0.000190
(2000,7000]	0.0522	5000	0.000010
(7000,100000]	0.0031	93000	0.000000



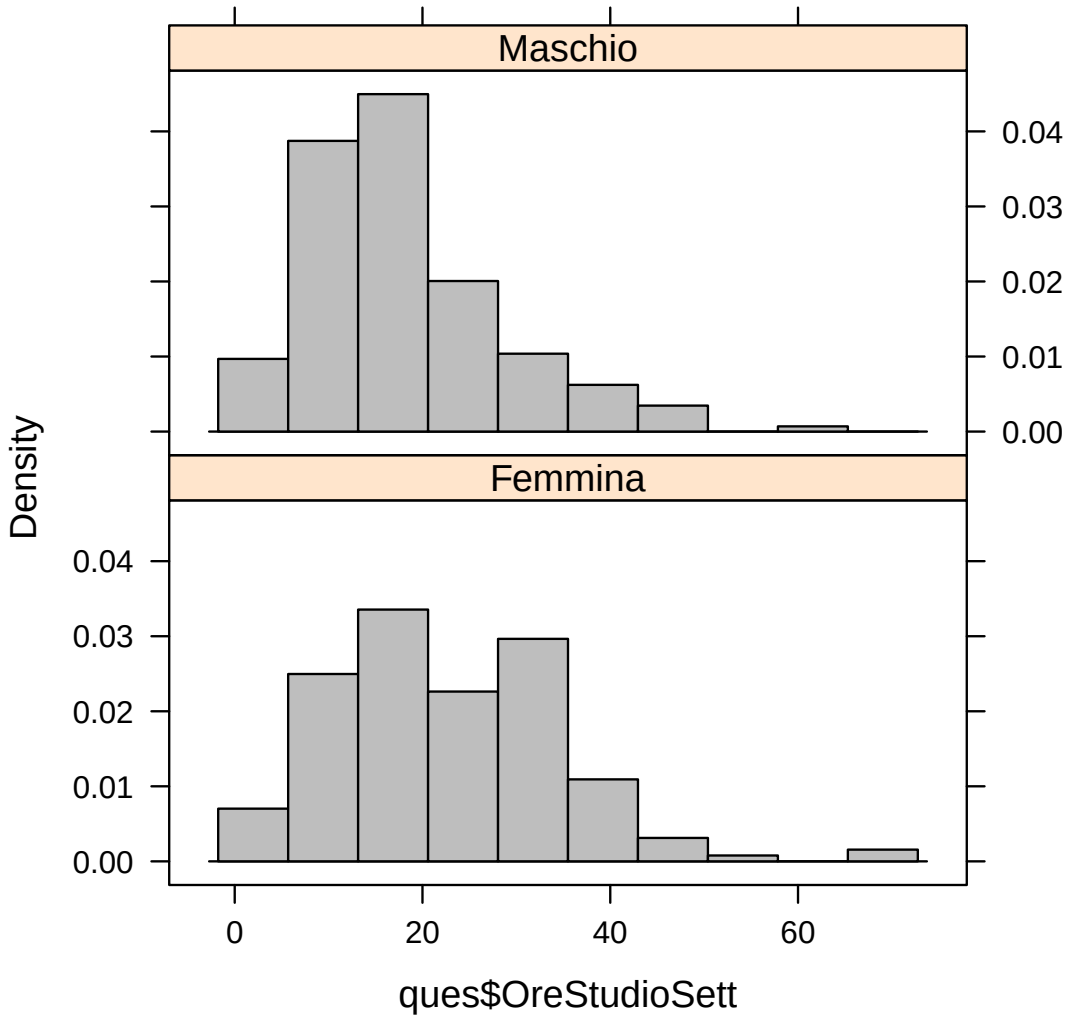
Vale anche per le distribuzioni condizionate

Esempio: altezze di maschi e femmine



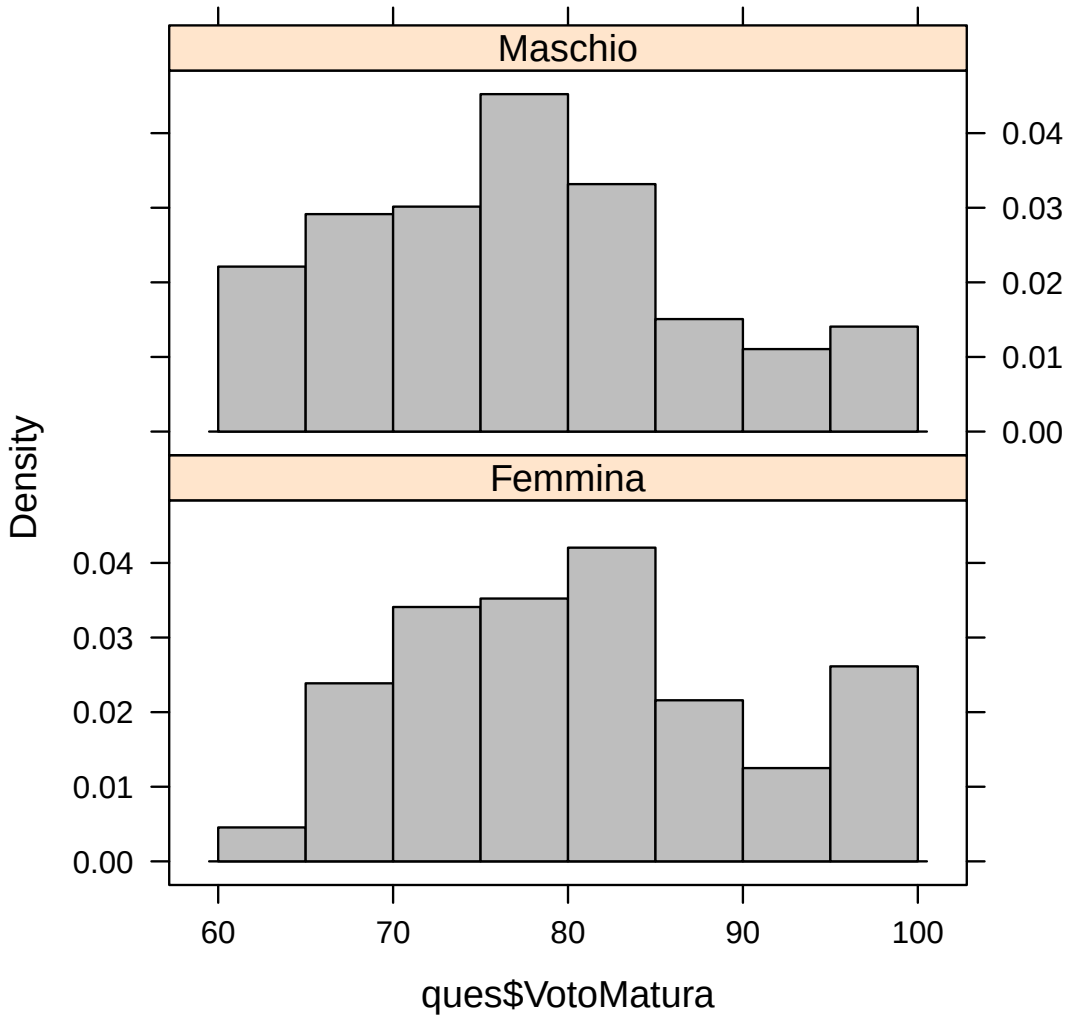
Vale anche per le distribuzioni condizionate

Esempio: ore di studio di maschi e femmine



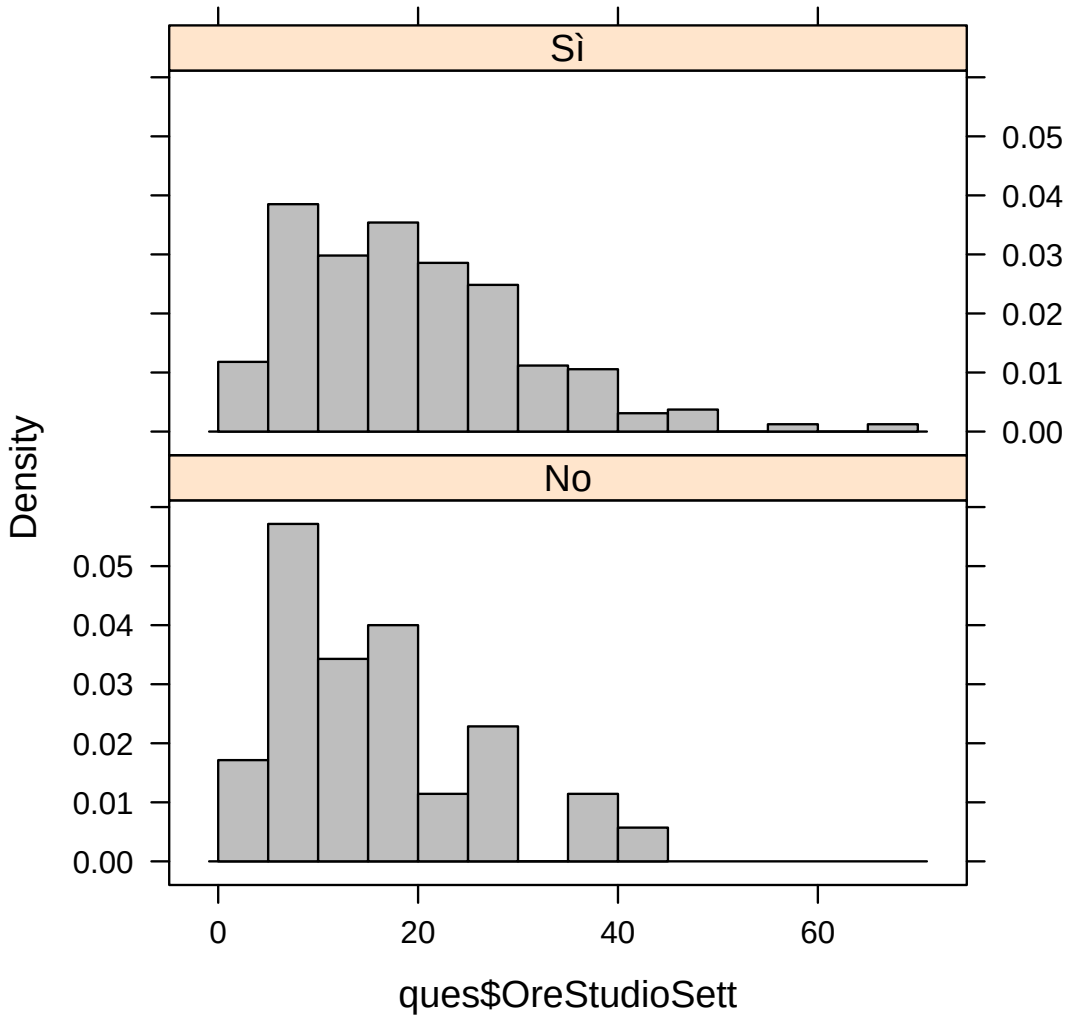
Vale anche per le distribuzioni condizionate

Esempio: voto di matura di maschi e femmine



Vale anche per le distribuzioni condizionate

Esempio: ore di studio settimanali - superamento Matematica



Terminologia

- Per variabili categoriali, la rappresentazione prende il nome di **diagramma a torta** o **diagramma a barre**.
- Per variabili discrete, la rappresentazione prende il nome di **diagramma a barre**.
- Per variabili continue, la rappresentazione prende il nome di **istogramma**.

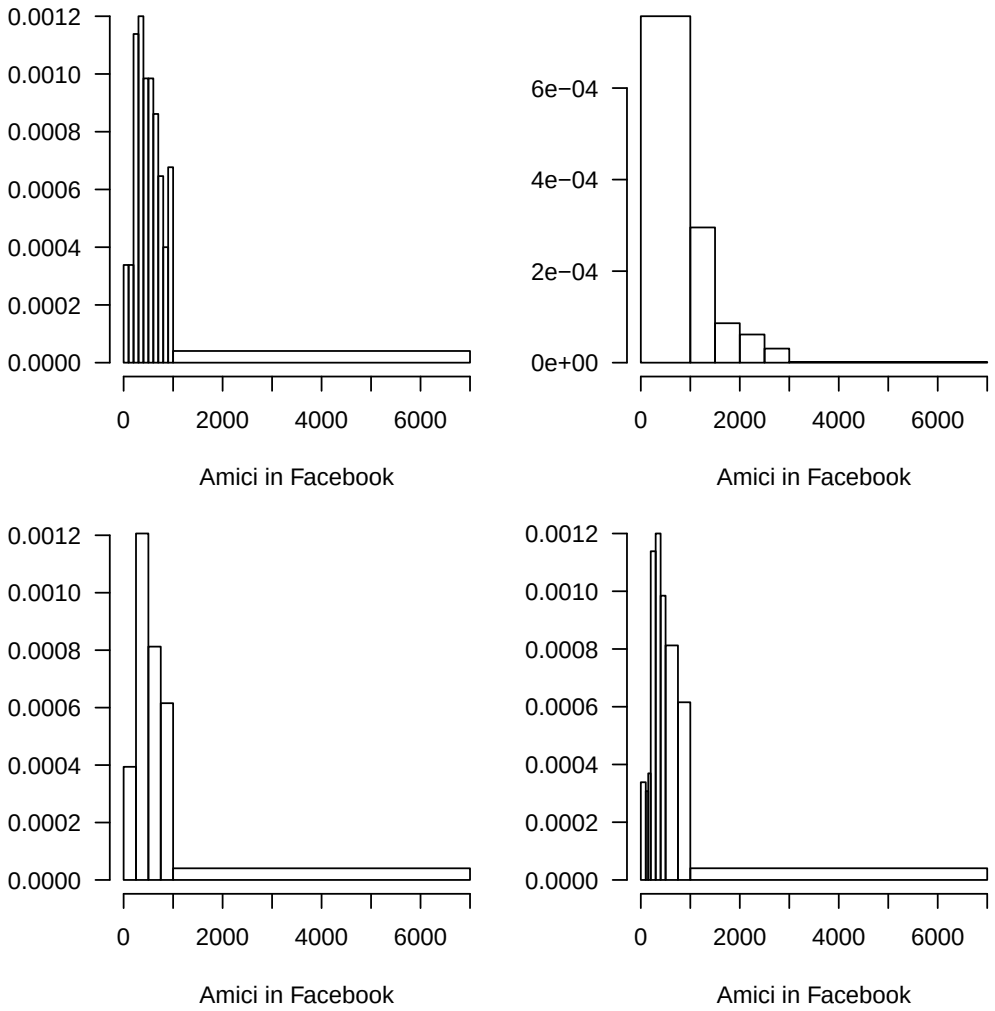
Osservazioni

Le rappresentazioni grafiche di distribuzioni di frequenza

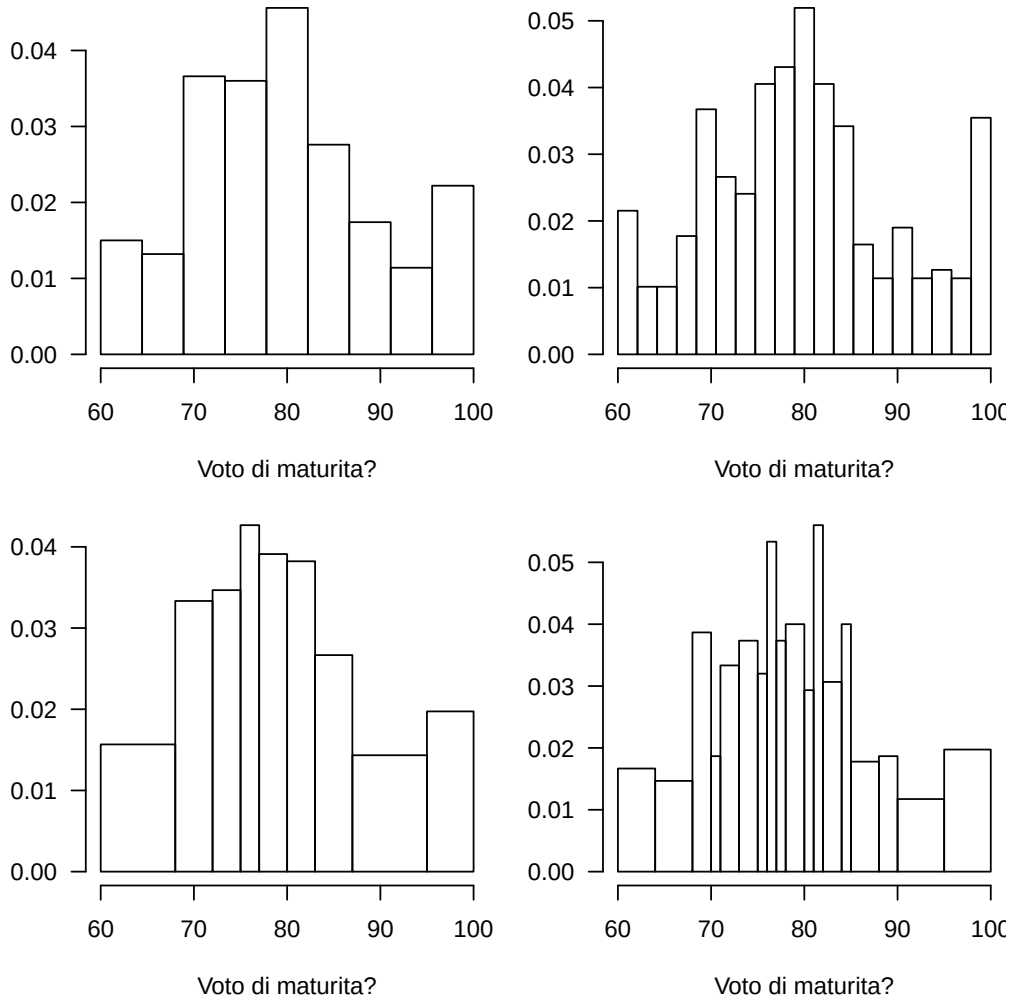
- forniscono una immagine della distribuzione dei dati: barre o scatole più alte rappresentano modalità più frequenti;
- aiutano a descrivere la **forma** della distribuzione dei dati;
- sono fortemente comunicative;
- ma devono essere ben costruite!

Esempio: amici di Facebook

Quale di questi istogrammi è utile? Quale fornisce troppi dettagli? Quale nasconde troppo?



Esempio: Voto di maturità?



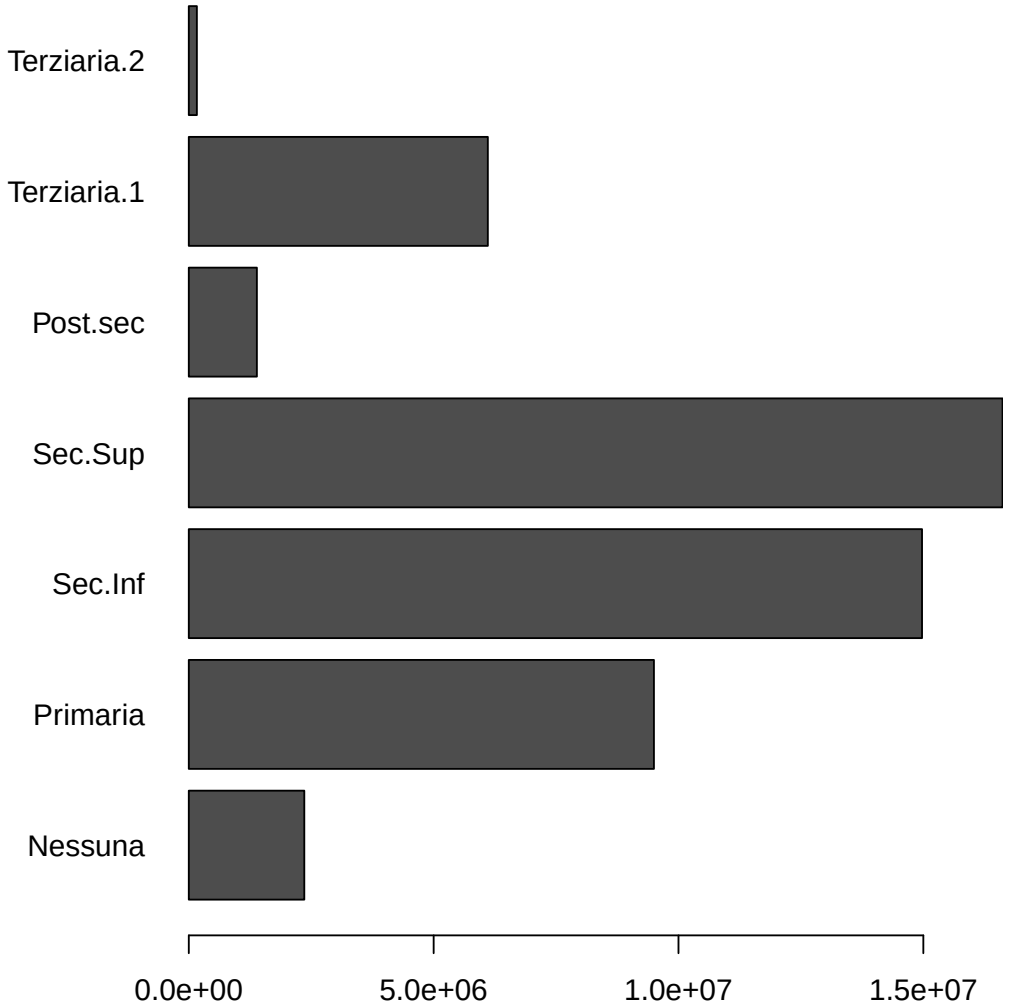
Osservazioni: ampiezza delle classi degli istogrammi (cont)

- Pochi intervalli, pochi dettagli.
- Troppi intervalli, troppi dettagli, probabilmente peculiari del campione a disposizione.
- È conveniente fare più di un grafico: provare differenti lunghezze per gli intervalli e poi scegliere.
- Il numero degli intervalli deve dipendere dal numero dei dati!

Esempio: titoli di studio

In base al censimento 2011, la distribuzione di frequenza assoluta del titolo di studio per la popolazione italiana sopra i 15 anni è

Titolo	Frequenza Italy
Nessuna	2358195.00
Primaria	9498111.00
Sec.Inf	14973047.00
Sec.Sup	16617577.00
Post.sec	1389813.00
Terziaria.1	6106337.00
Terziaria.2	164622.00



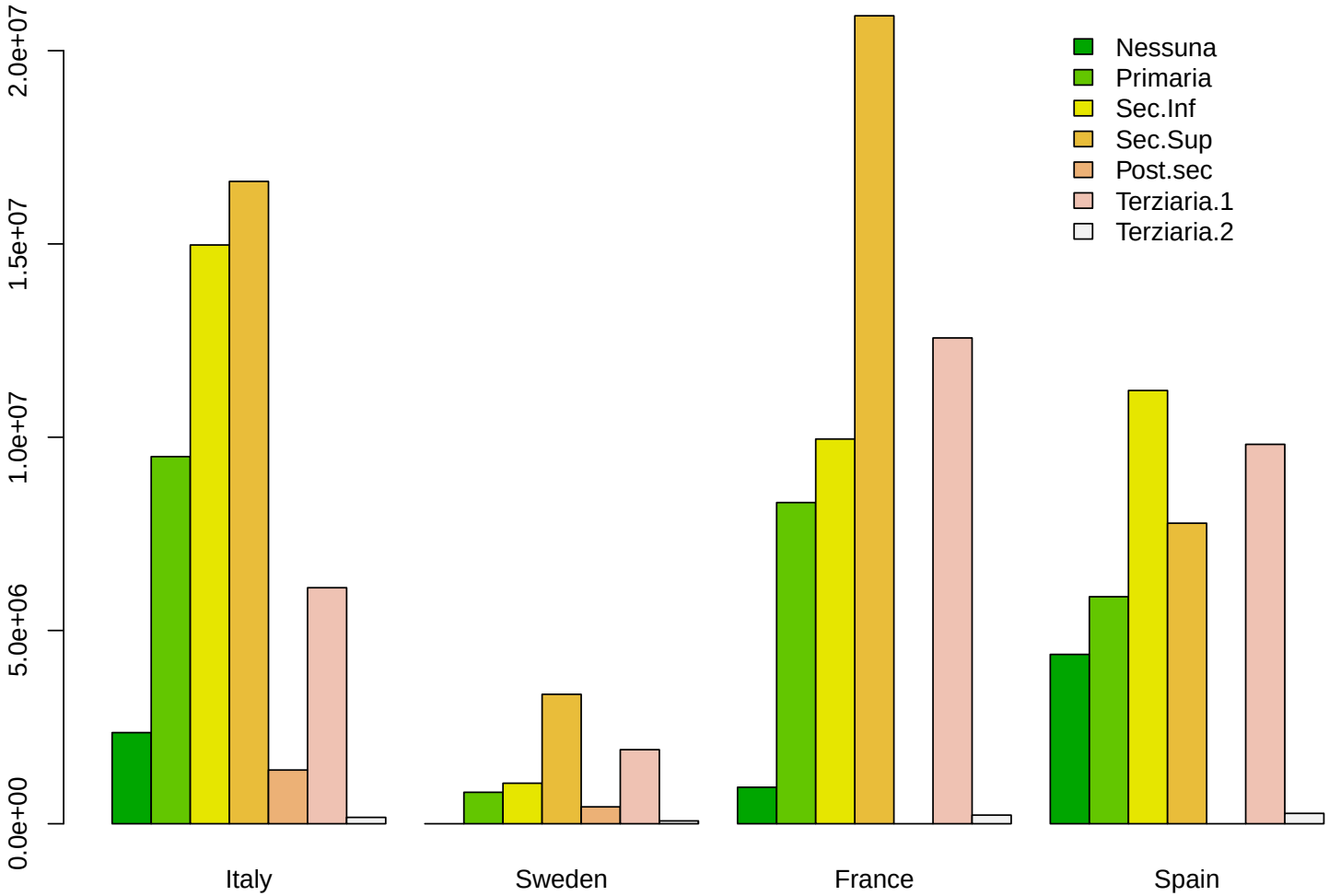
Titoli di studio

Potrei essere interessato a confronti tra paesi

Titolo	Frequenza			
	Italy	Sweden	France	Spain
Nessuna	2358195	0	943760	4381305
Primaria	9498111	813963	8309752	5873435
Sec.Inf	14973047	1048392	9952880	11208975
Sec.Sup	16617577	3348458	20902738	7775165
Post.sec	1389813	436909	0	0
Terziaria.1	6106337	1916300	12567952	9815605
Terziaria.2	164622	73725	223497	268390

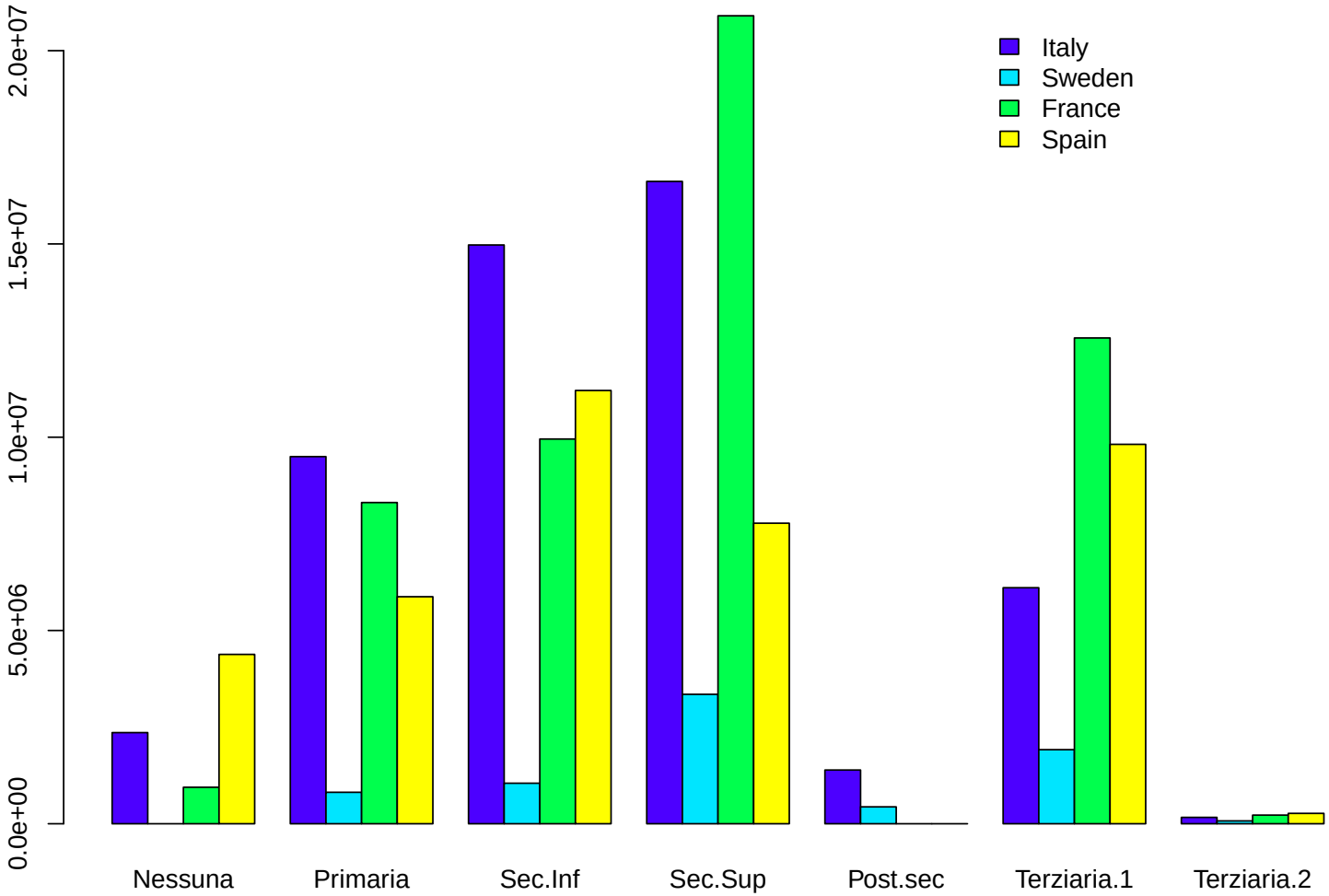
Titoli di studio

Potrei essere interessato a confronti tra paesi



Titoli di studio

Potrei essere interessato a confronti tra paesi



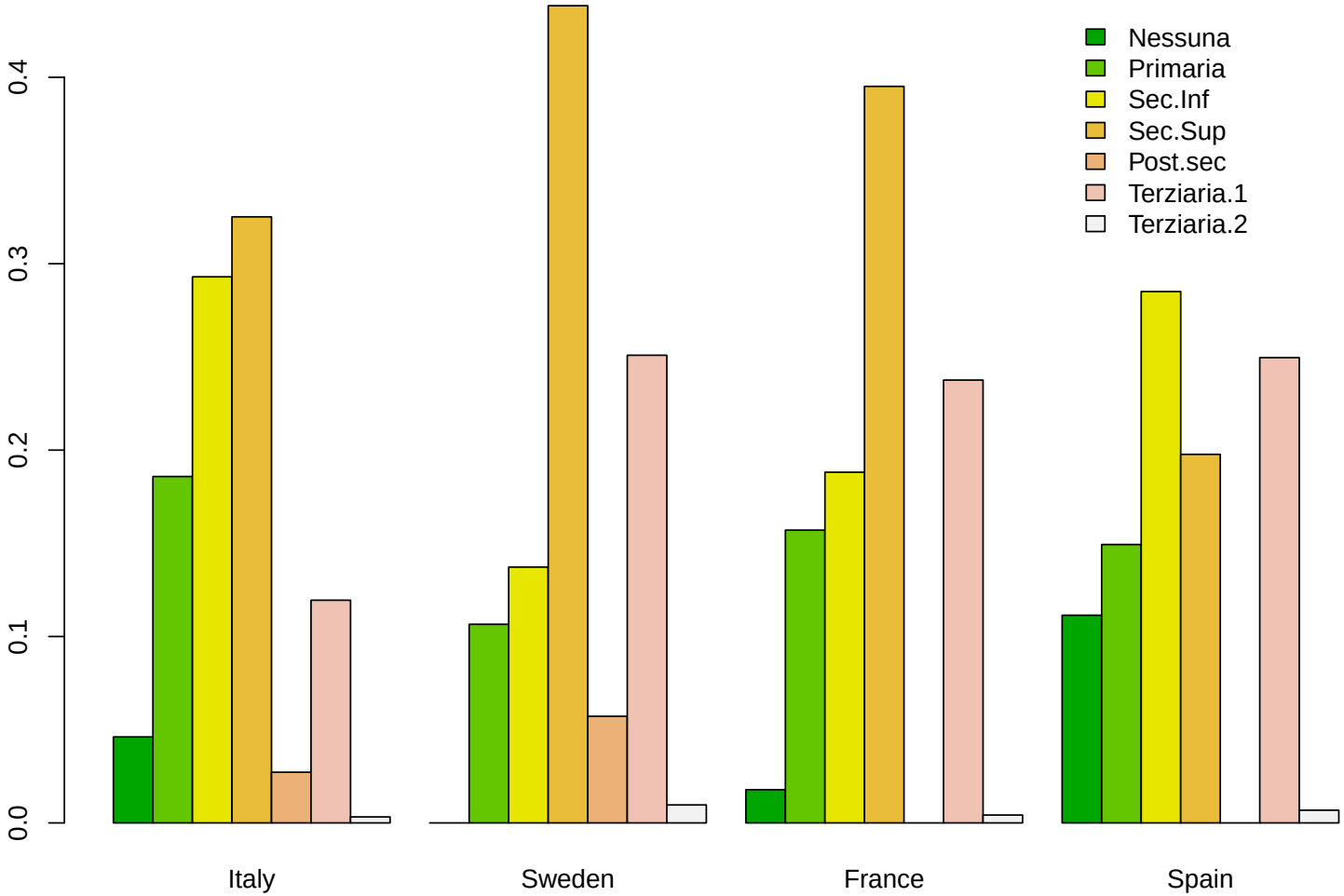
Titoli di studio

Per rendere possibile il confronto si passa alle frequenze relative

Titolo	Frequenza			
	Italy	Sweden	France	Spain
Nessuna	0.0461	0.0000	0.0178	0.1114
Primaria	0.1858	0.1066	0.1571	0.1494
Sec.Inf	0.2930	0.1373	0.1881	0.2850
Sec.Sup	0.3251	0.4384	0.3951	0.1977
Post.sec	0.0272	0.0572	0.0000	0.0000
Terziaria.1	0.1195	0.2509	0.2376	0.2496
Terziaria.2	0.0032	0.0097	0.0042	0.0068

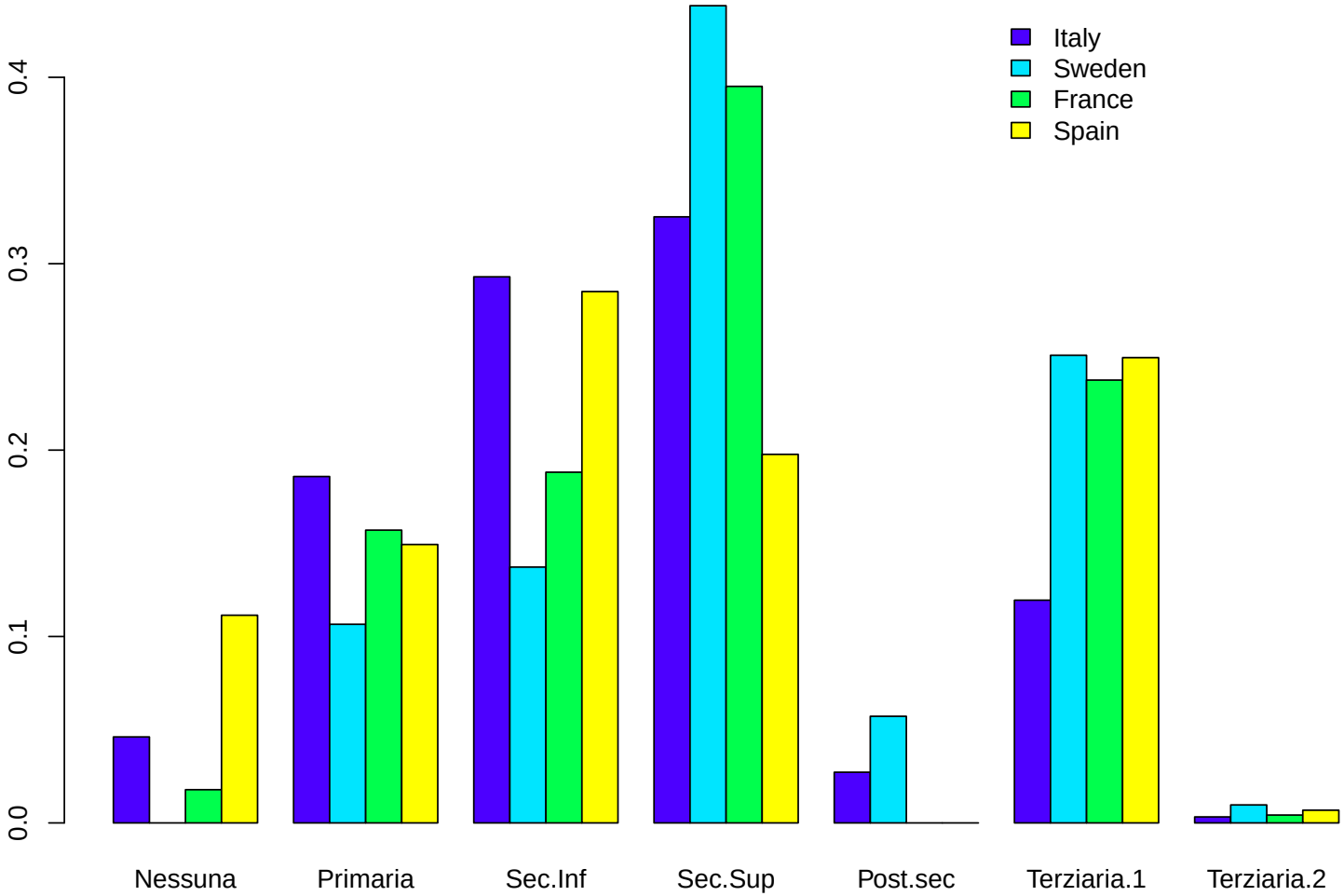
Titoli di studio

Per rendere possibile il confronto si passa alle frequenze relative



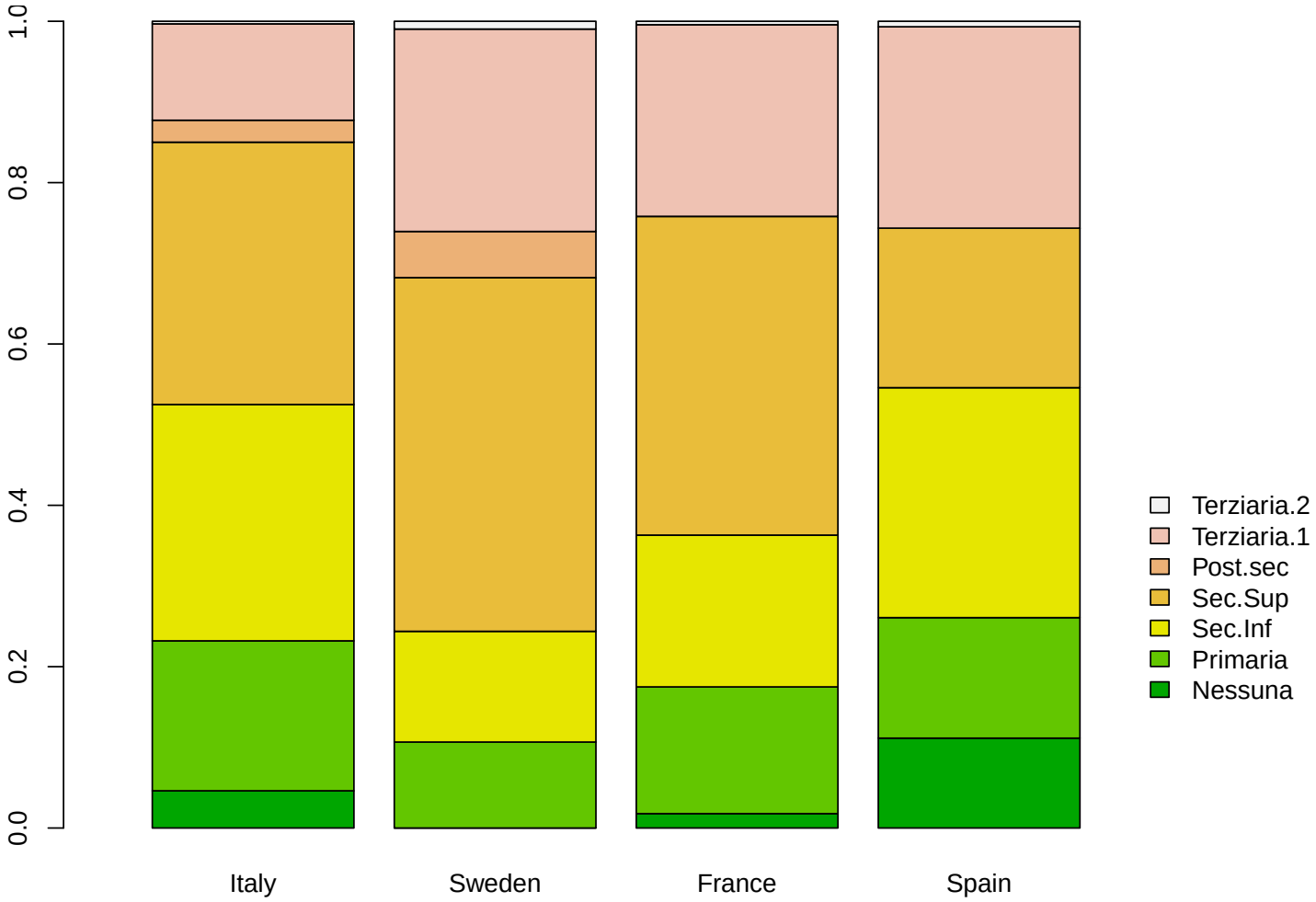
Titoli di studio

Per rendere possibile il confronto si passa alle frequenze relative



Titoli di studio

Per rendere possibile il confronto si passa alle frequenze relative



Titoli di studio, osservazioni

- Se la domanda è "in quali paesi si studia di più", le frequenze assolute non consentono un confronto agevole perché le popolazioni di riferimento sono molto diverse.
- Si passa allora alle frequenze relative per ciascun paese.
- Il confronto può essere fatto affiancando dei diagrammi a barre, il modo in cui le si affianca mette in evidenza cose diverse.
- Le barre possono anche essere sovrapposte per mettere in luce le diverse composizioni delle popolazioni.

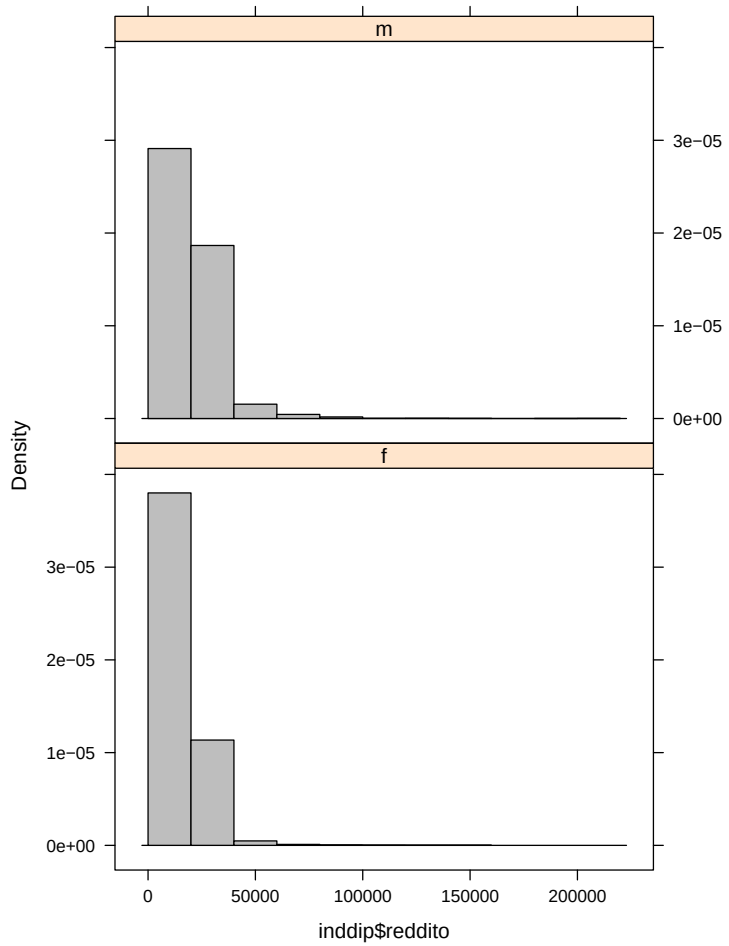
Reddito degli italiani (lav. dipendenti) e genere

Consideriamo i dati dell'indagine istat sulle famiglie, in particolare i dati individuali sui lavoratori dipendenti.

Consideriamo il reddito condizionatamente al sesso.

—

Il grafico non dice molto, questo perché ci sono alcuni redditi molto elevati e i restanti sono schiacciati.



Reddito degli italiani (lav. dipendenti) e genere

Consideriamo i dati dell'indagine istat sulle famiglie, in particolare i dati individuali sui lavoratori dipendenti.

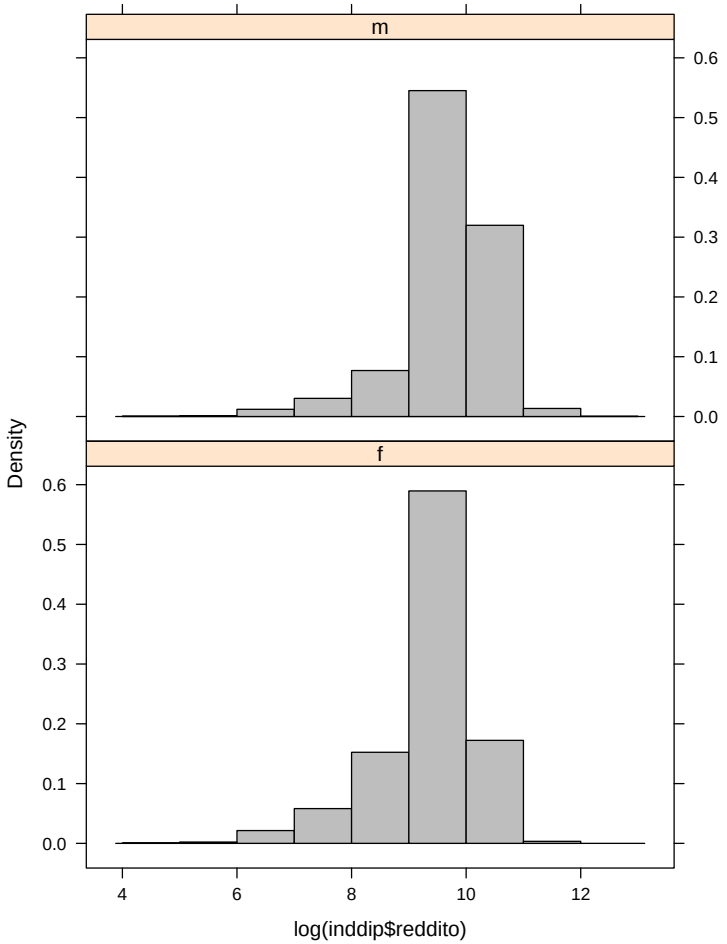
Consideriamo il reddito condizionatamente al sesso.

—

Il grafico non dice molto, questo perché ci sono alcuni redditi molto elevati e i restanti sono schiacciati.

—

Conviene allora passare ai logaritmi dei redditi, per esaminare meglio il centro della distribuzione.



Reddito degli italiani (lav. dipendenti) e genere

Consideriamo i dati dell'indagine istat sulle famiglie, in particolare i dati individuali sui lavoratori dipendenti.

Consideriamo il reddito condizionatamente al sesso.

—

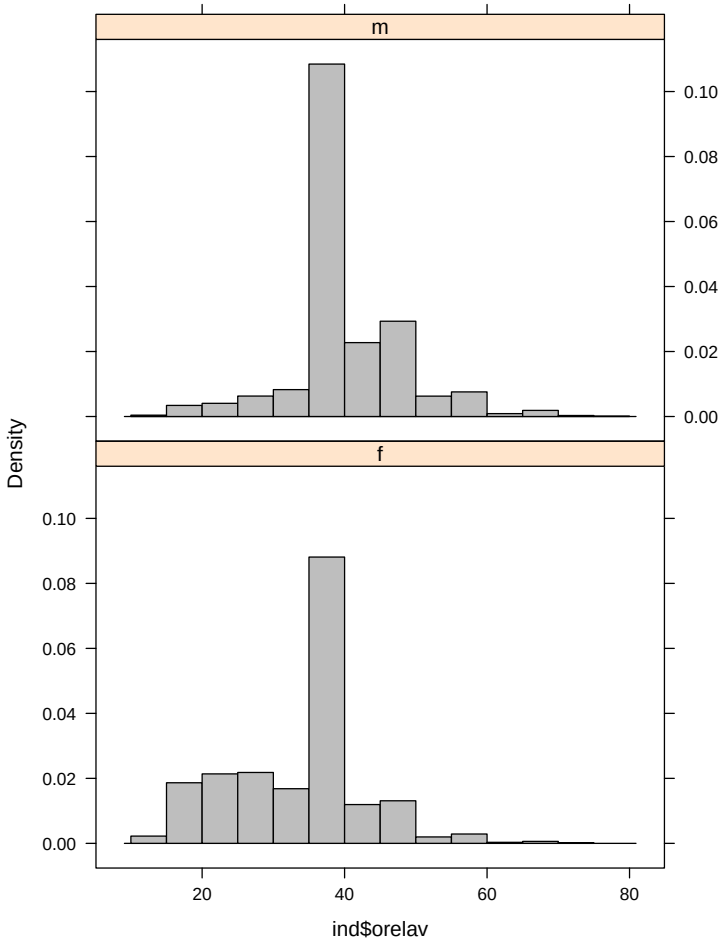
Il grafico non dice molto, questo perché ci sono alcuni redditi molto elevati e i restanti sono schiacciati.

—

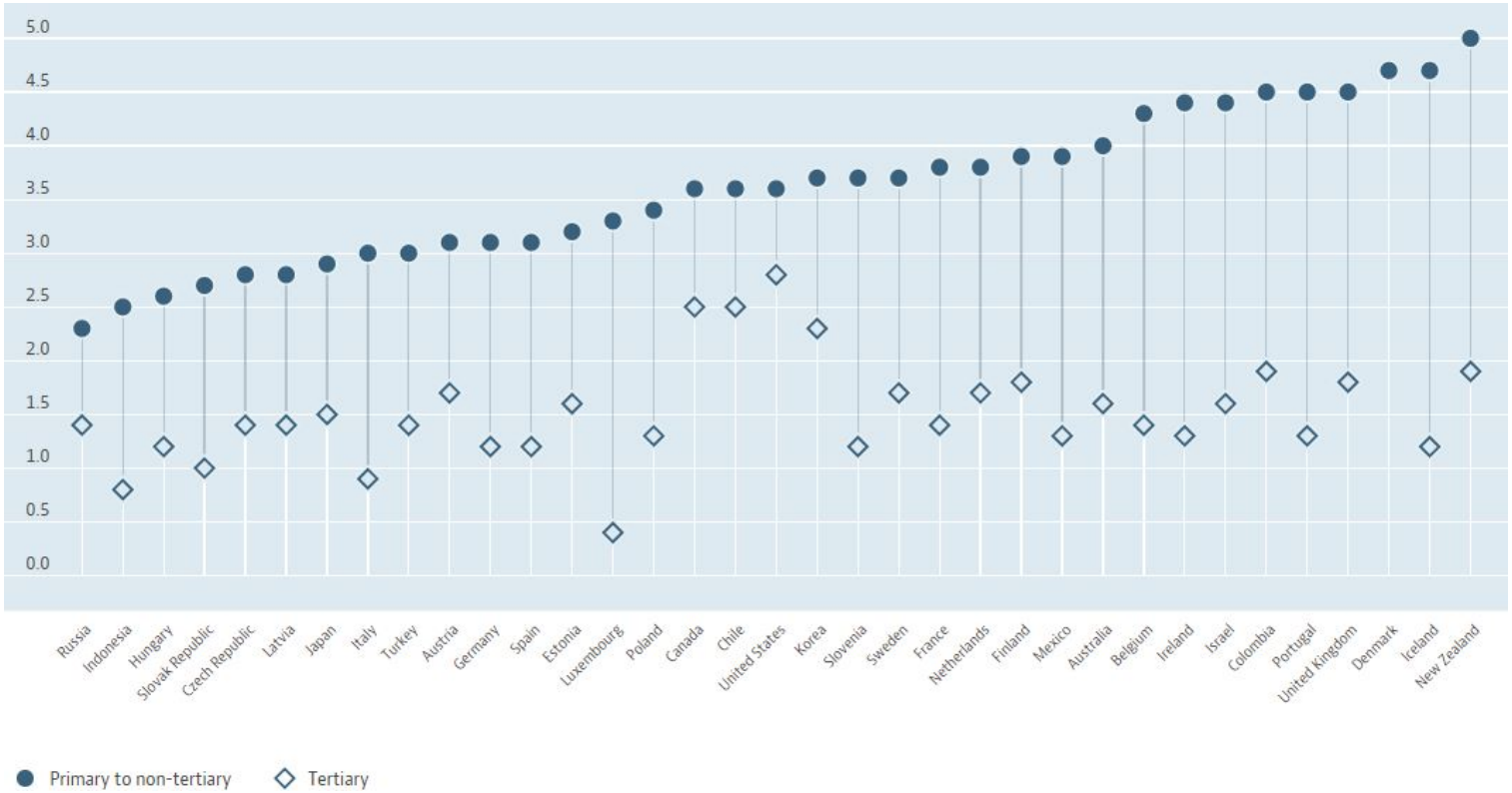
Conviene allora passare ai logaritmi dei redditi, per esaminare meglio il centro della distribuzione.

—

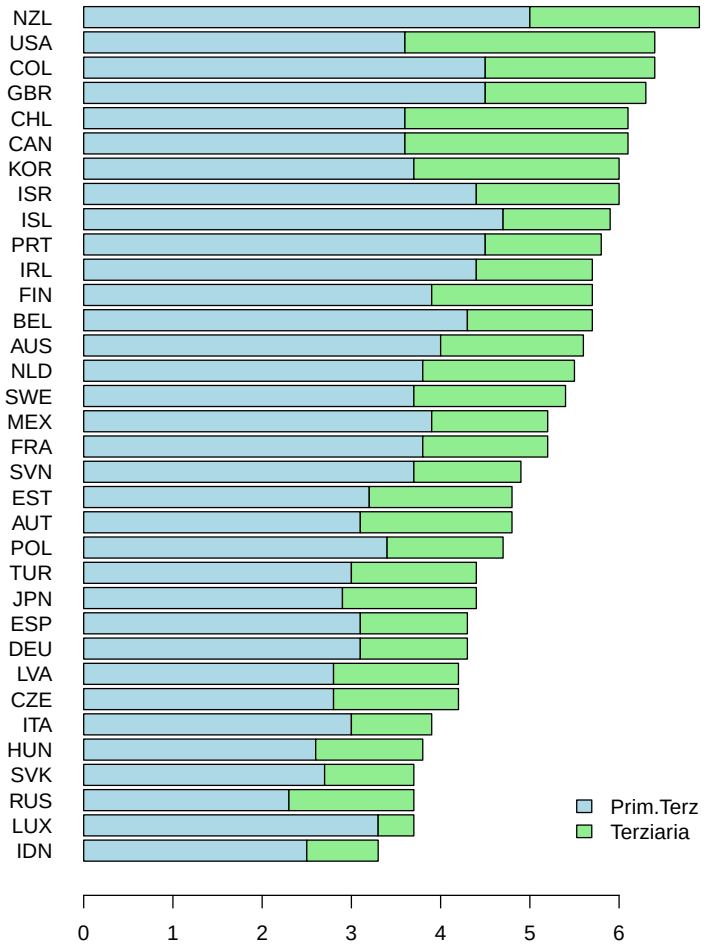
Mettiamo in relazione quanto visto con le ore lavorate.



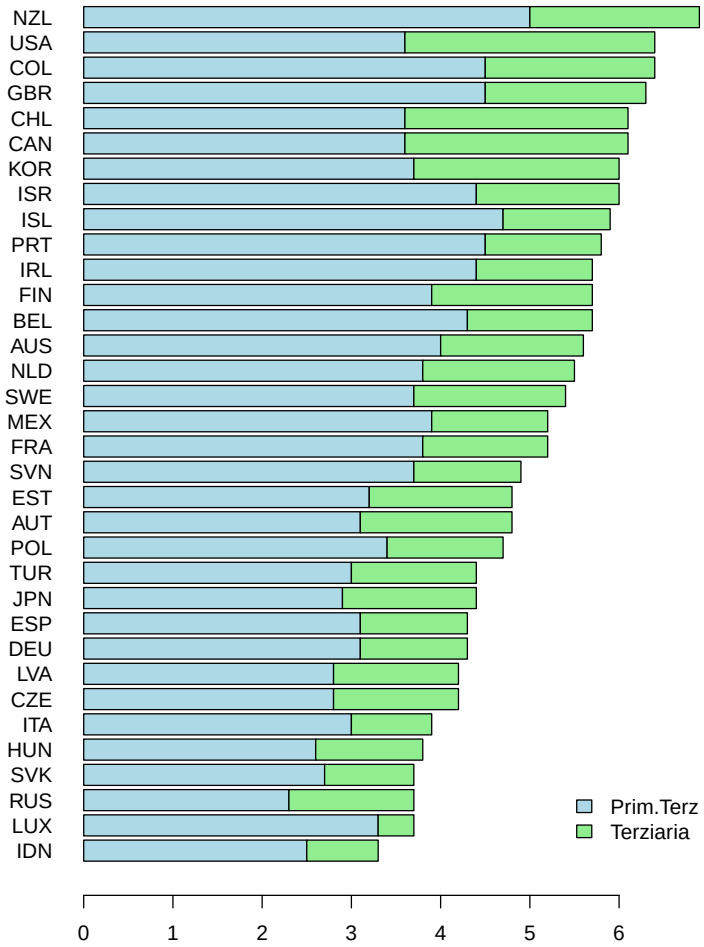
Spesa per l'istruzione (% del PIL)

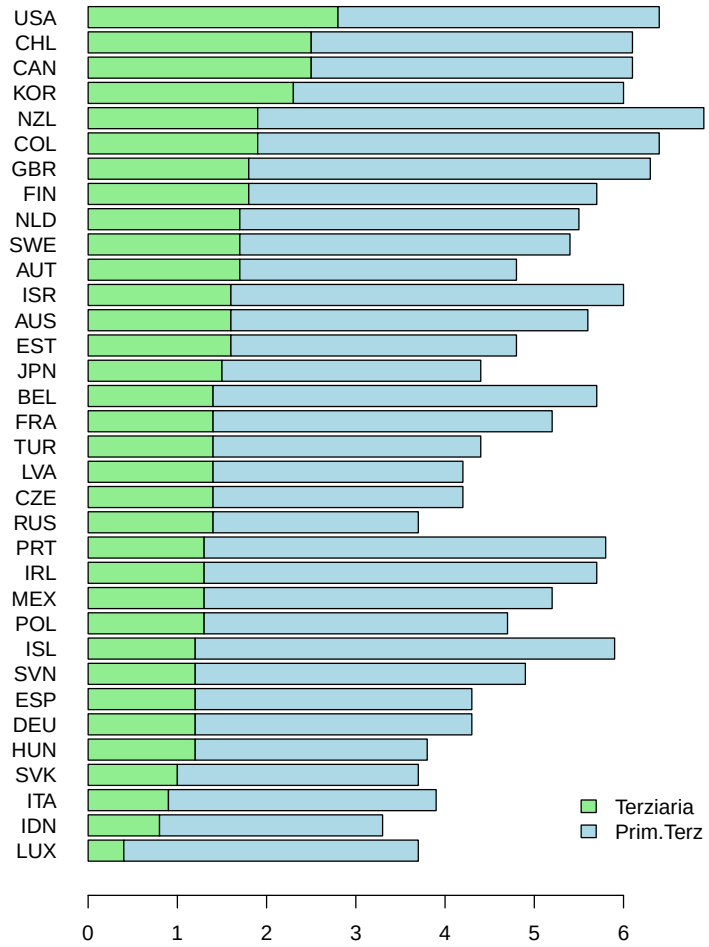


Spesa per l'istruzione (% del PIL)



Spesa per l'istruzione (% del PIL)





Spesa per l'istruzione (% del PIL), osservazioni

- Il fatto che la spesa sia espressa in percentuale del PIL rende le osservazioni dei vari paesi confrontabili (se fosse la spesa in euro dovremmo relativizzare).
- Usando diversi ordinamenti si mettono in evidenza aspetti diversi.