Misure di Frequenze e Rappresentazione Grafica dei Dati

Contents

In questa lezione, inizieremo a vedere dei primi strumenti per riassumere le caratteristiche fondamentali dei dati.

10.1. Frequenze Assolute e Relative

10.1.1. Frequenze Assolute

Un primo modo di descrivere i dati consiste nel calcolare il numero di volte in cui ciascun valore appare. Queste sono chiamate “frequenze assolute”. Le frequenze assolute sono in genere calcolate per variabili discrete in cui le osservazioni assumono un numero finito di valori.

Siano

$$a_1, a_2, \ldots, a_3$$

i valori che la variabile in considerazione può assumere.

Le frequenze assolute n_i sono definite come il numero di volte che a_i appare nel campione. Si noti che:

$$\sum_i n_i = n$$

Dove n è il numero totale di elementi nel campione.

10.1.1.1. Esempi

Consideriamo un semplice campione di 10 pazienti per i quali sono stati rilevati dei dati. Consideriamo una variabile `gender` che indica il genere dei pazienti:

```
0    M
1    F
2    M
3    M
4    M
5    F
6    F
7    F
8    F
9    F
dtype: object
```

I valori univoci in questo semplici esempi saranno due:

```
0    M
1    F
dtype: object
```

Le frequenze assolute del campione in oggetto sono riassunte nella tabella seguente:

```
F    6
M    4
dtype: int64
```

Consideriamo un dataset un po' più complesso, contenente pesi (in libbre), altezze (in pollici) e sesso di diversi soggetti. Il dataset avrà il seguente aspetto:

	sex	height	weight
0	M	74	53.484771
1	M	70	38.056472
2	F	61	34.970812
3	M	68	35.999365
4	F	66	34.559390
...	...	...	...
4226	F	69	23.862436
4227	M	69	38.262182
4228	F	64	34.970812
4229	F	64	28.388071
4230	F	61	22.628172

4231 rows × 3 columns

Notiamo che i valori delle altezze sono quantizzate. I valori univoci in questo caso saranno i seguenti:

```
array([74, 70, 61, 68, 66, 65, 64, 67, 72, 71, 76, 69, 63, 75, 60, 59
```

Le frequenze assolute in questo caso saranno le seguenti:

```
59     52
60    130
61    411
63    291
64    435
65    351
66    391
67    377
68    355
69    302
70    272
71    235
72    260
73    146
74    104
75     72
76     47
Name: height, dtype: int64
```

10.1.1.2. Diagramma a Barre delle Frequenze Assolute

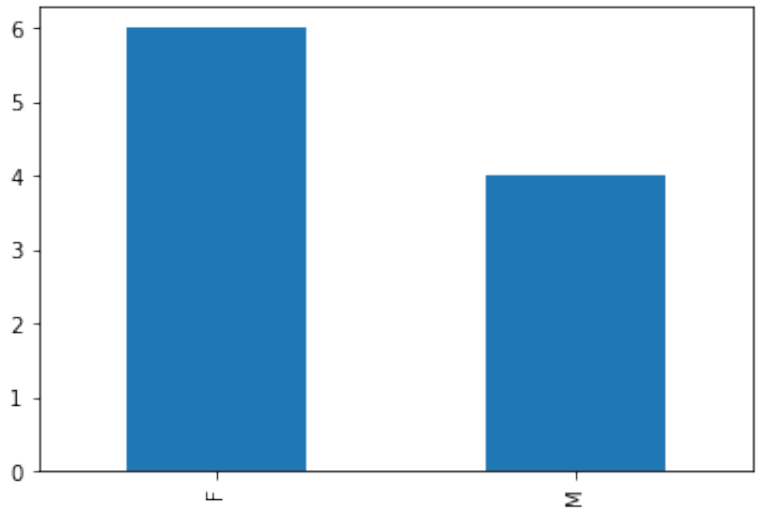
Le frequenze dei dati possono essere rappresentate graficamente mediante un diagramma a barre che pone sull'asse delle x il valore univoco (a_i) e che rappresenta la frequenza assoluta n_i come altezza della barra.

10.1.1.2.1. Esempi

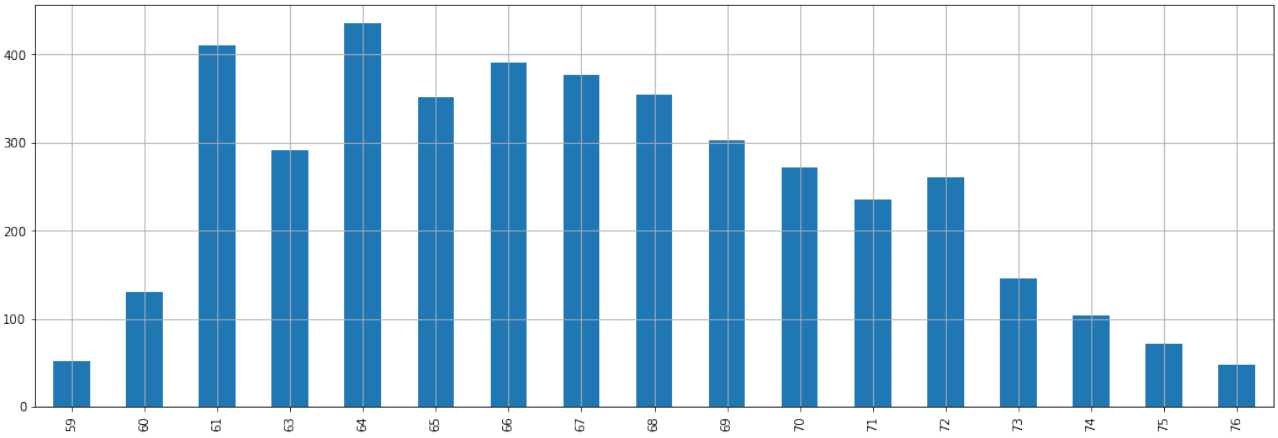
Nel caso del campione `gender`:

```
data.value_counts().plot.bar()
```

```
<AxesSubplot:>
```



Il diagramma a barre delle frequenze assolute delle altezze nel nostro campione di pesi e altezze è il seguente:



Il grafico sopra ci dice qualcosa sul numero di occorrenze di un dato valore e ci da anche una idea di quali valori siano più o meno frequenti. Vediamo che i dati seguono una forma “a campana”. Troveremo spesso questo tipo di forma e ne parleremo meglio in seguito.

10.1.2. Frequenze Relative (Probability Mass Function)

Le frequenze assolute ci permettono di farci un’idea più precisa su come i dati sono distribuiti, indipendentemente dalla dimensione del nostro campione. Sappiamo ad esempio che il campione contiene più individui di altezza pari a $167.64\ cm$ che individui di altezza pari a $193.04\ cm$. Tuttavia, tale rappresentazione è legata al numero totale di elementi contenuti nel campione. Ad

esempio, un campione distribuito in maniera simile, ma con più osservazioni, darà luogo a frequenze assolute più grandi. Possiamo ottenere una rappresentazione indipendente rispetto alla dimensione del campione mediante l'analisi delle frequenze relative, definite come seguono:

$$f_j = f(a_j) = \frac{n_j}{n}, j = 1, 2, \ldots, k$$

Si noti che, vista la definizione, si avrà:

$$n_j \leq n \Rightarrow f_j \leq 1 \; \forall j$$

$$\sum_j f_j = \sum_j \frac{n_j}{n} = \frac{1}{n} \sum_j n_j = \frac{n}{n} = 1$$

10.1.2.1. Esempi

Nel caso del nostro piccolo campione `gender` avremo:

```
data.value_counts(normalize=True)
```

```
F    0.6
M    0.4
dtype: float64
```

Nel caso del nostro dataset di altezze, avremo:

```
59    0.012290
60    0.030726
61    0.097140
63    0.068778
64    0.102813
65    0.082959
66    0.092413
67    0.089104
68    0.083905
69    0.071378
70    0.064287
71    0.055542
72    0.061451
73    0.034507
74    0.024580
75    0.017017
76    0.011108
Name: height, dtype: float64
```

È possibile verificare che tutti i numeri sono compresi tra zero e uno e che la somma dei valori è pari a 1.

Come abbiamo visto, questa rappresentazione è nota anche come **probability mass function (PMF)** e associa ad **ogni valore discreto** presente nel campione una **probabilità**. Possiamo ad esempio dire che la probabilità di trovare un individuo di altezza $67 inches$ (circa $170.18 cm$) è pari a 0.089104.

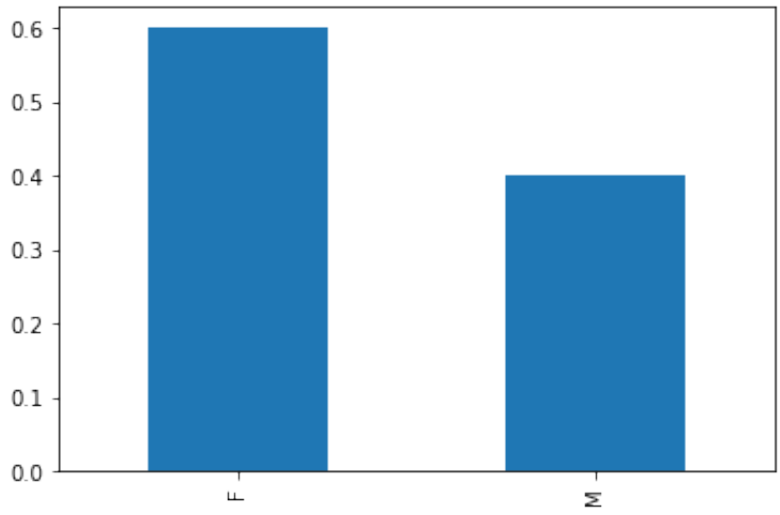
10.1.2.2. Diagramma a Barre delle Frequenze Relative

Plottando le frequenze relative con un diagramma a barre, otteniamo un grafico molto simile a quello delle frequenze relative, ma con una differente scala sull'asse delle y.

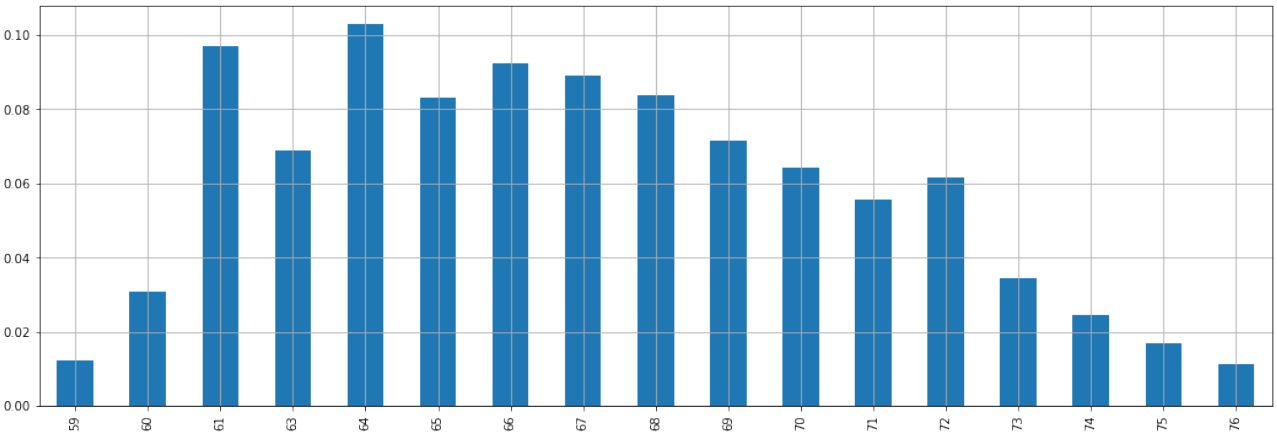
10.1.2.2.1. Esempi

Nel caso del campione `gender` avremo:

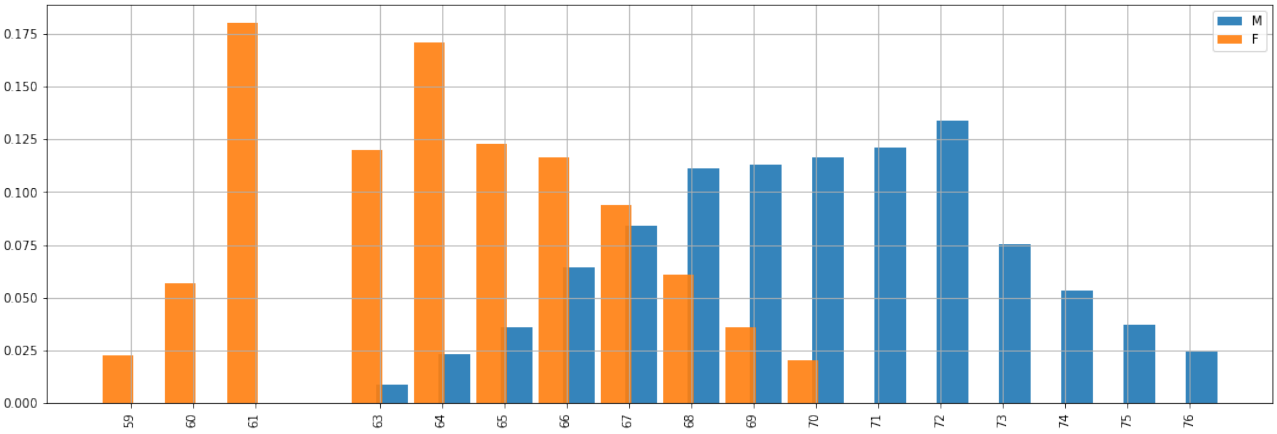
```
<AxesSubplot:>
```



Nel caso del campione pesi-altezze, avremo:



I diagrammi a barre delle frequenze relative sono utili per confrontare tra di loro campioni diversi. Ad esempio, possiamo considerare le altezze di uomini e donne nel dataset pesi-altezze:



Dal confronto sopra, possiamo già fare delle considerazioni qualitative sui due campioni. In particolare notiamo (poco sorprendentemente) che gli uomini sono

generalmente più alti delle donne. Ciò non vuol dire che non esistano uomini più bassi di alcune donne o viceversa, ma ragionevolmente si tratta di casi meno frequenti.

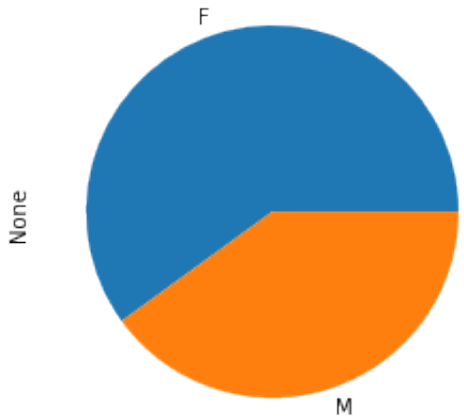
10.1.2.3. Grafici a Torta

In alternativa ai grafici a barre, le frequenze relative possono essere visualizzate anche mediante dei grafici a torta.

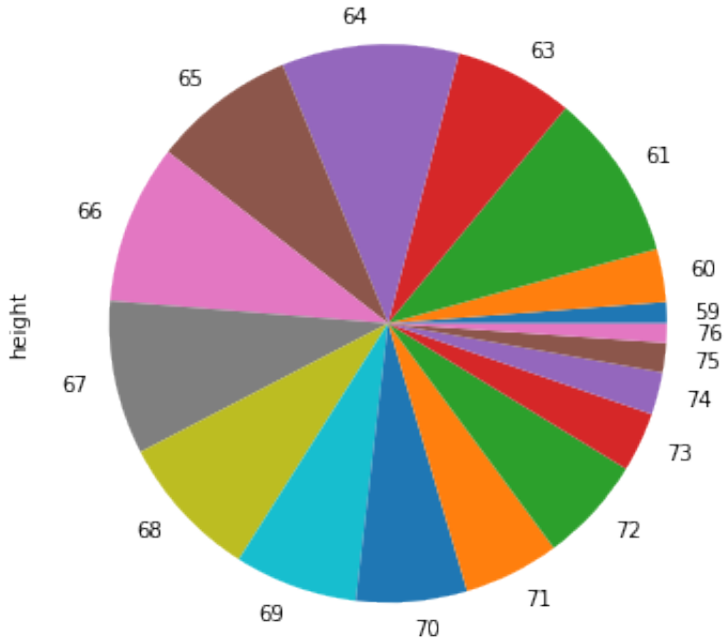
10.1.2.3.1. Esempi

Nel caso del campione `gender`:

```
<AxesSubplot:ylabel='None'>
```



Nel caso del campione pesi-altezze:



10.2. Empirical Cumulative Distribution Function (ECDF)

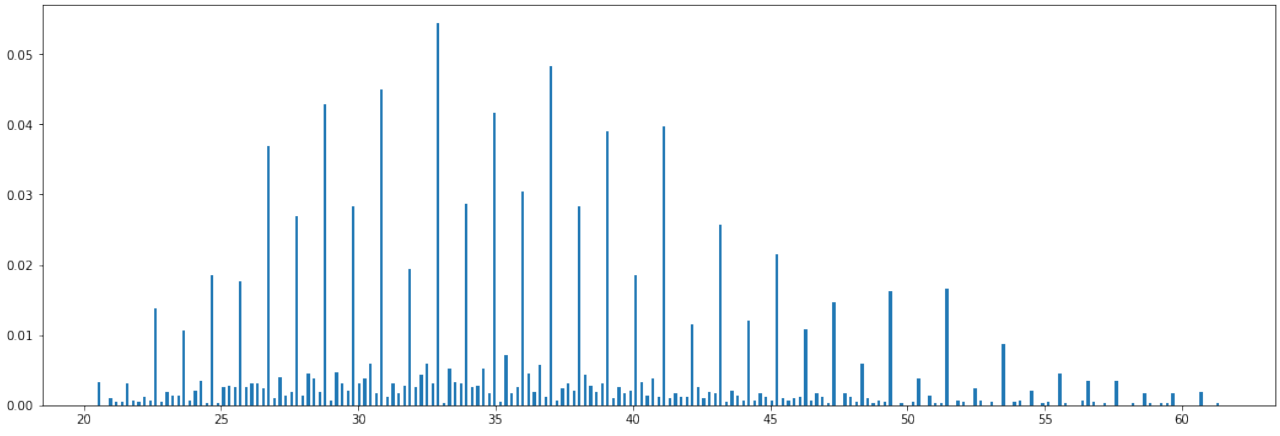
Le frequenze relative funzionano particolarmente bene quando i valori unici sono pochi. Quando invece il numero di valori univoci cresce, le frequenze discrete calcolate per i valori diventano molto piccole e dunque soggette a rumore (ad esempio dovuto ad errori di misura).

Torniamo al nostro esempio di dataset di pesi e altezze:

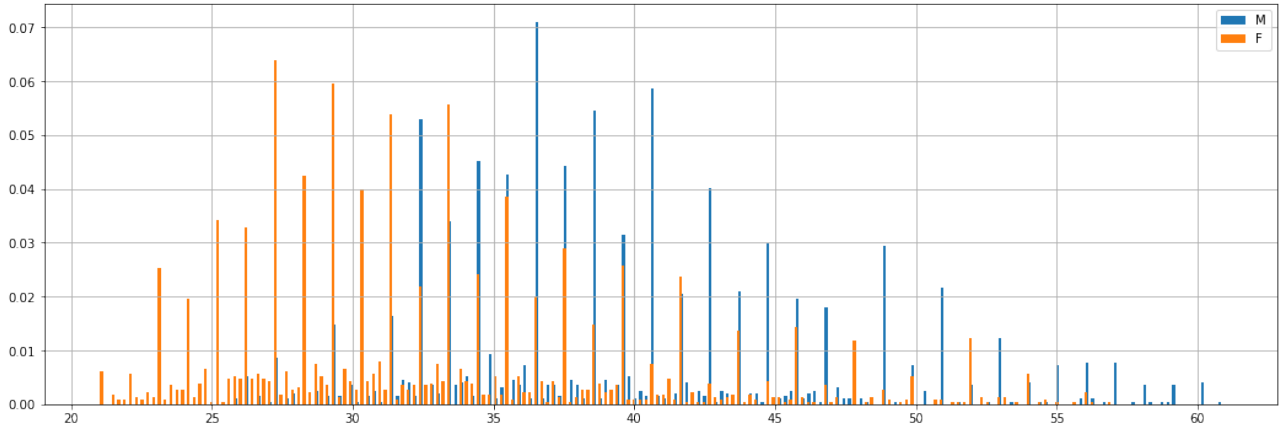
	sex	height	weight
0	M	74	53.484771
1	M	70	38.056472
2	F	61	34.970812
3	M	68	35.999365
4	F	66	34.559390
...	...	...	...
4226	F	69	23.862436
4227	M	69	38.262182
4228	F	64	34.970812
4229	F	64	28.388071
4230	F	61	22.628172

4231 rows × 3 columns

I valori dei pesi, a differenza di quelli delle altezze, non sono quantizzati. Un grafico a barre delle frequenze relative (o Probability Mass Function) avrebbe questo aspetto:



Come possiamo notare, abbiamo ottenuto una rappresentazione grafica più “rumorosa”. Supponiamo adesso di voler confrontare le distribuzioni dei pesi di donne e uomini:



Le due rappresentazioni sono affette da rumore dovuto alla natura discreta dei dati per cui due valori molto vicini vengono trattati come due casi distinti nel calcolo delle probabilità. Vedremo alcuni modi per ovviare a questo problema. Uno di essi consiste nel calcolare una **Empirical Cumulative Distribution Function (ECDF)**. Una **ECDF** calcola per un valore x la somma delle frequenze relative di tutti i valori y minori o uguali a x:

$$ECDF(x) = \sum_{a_j : a_j \leq x} f(a_j)$$

Dove a_j sono i valori univoci all'interno del campione, mentre in generale $x \in \mathbb{R}$.

Consideriamo un semplice dataset di valori numerici:

```
0     1
1     5
2     2
3     6
4     5
5     4
6     3
7     5
8     4
9     2
10    4
11    5
12    6
13    4
14    4
15    3
dtype: int64
```

Le frequenze relative saranno le seguenti:

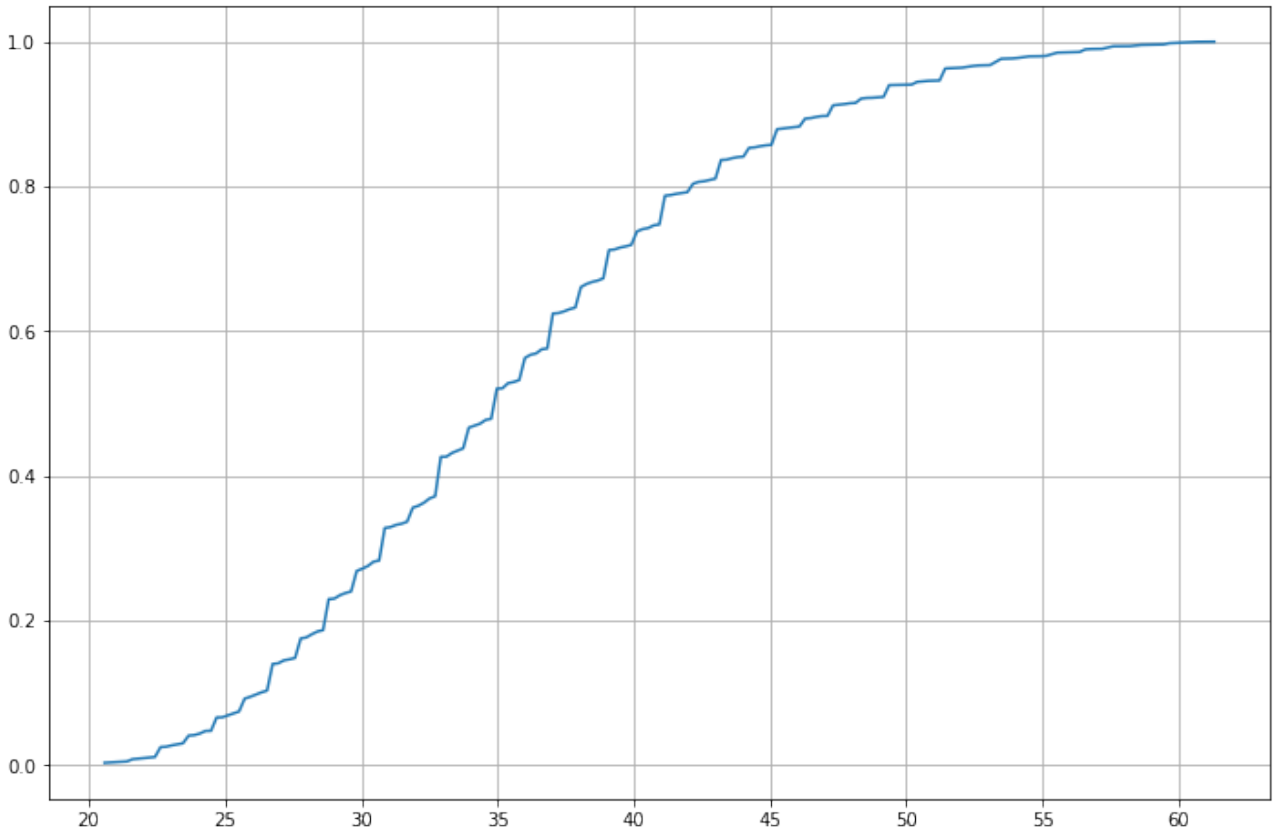
```
1    0.0625
2    0.1250
3    0.1250
4    0.3125
5    0.2500
6    0.1250
dtype: float64
```

La ECDF calcolata sui valori univoci sarà la seguente:

```
1    0.0625
2    0.1875
3    0.3125
4    0.6250
5    0.8750
6    1.0000
dtype: float64
```

Da notare che i valori della ECDF sono sempre crescenti e l'ultimo valore è pari a 1 (la somma di tutte le frequenze relative).

Una ECDF può essere rappresentata graficamente mettendo i valori di x sulle ascisse e i valori di $ECDF(x)$ sulle ordinate. Ad esempio, la ECDF dei pesi nel nostro dataset di pesi-altezze sarà la seguente:

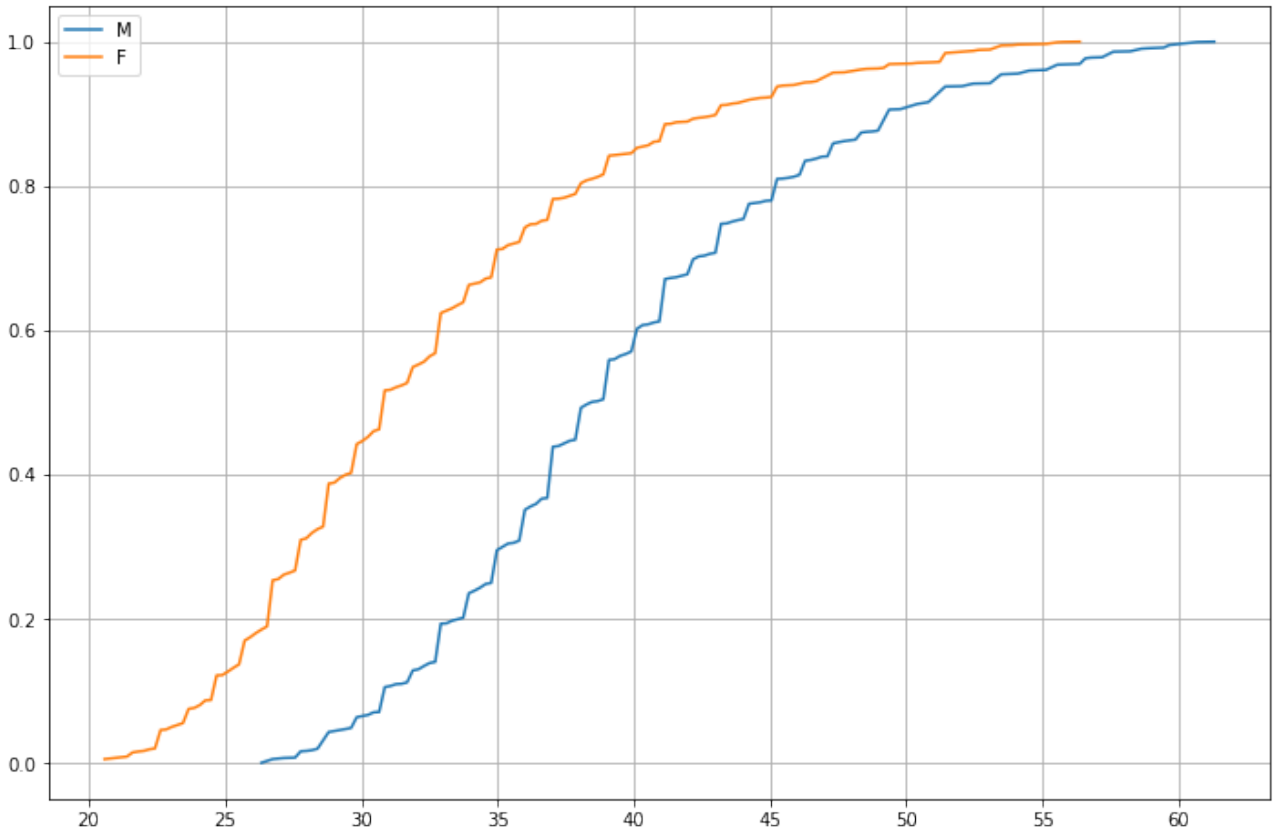


La definizione della ECDF ci dice che, dato un punto di coordinate (x, y), la somma delle frequenze degli elementi minori o uguali a x è pari a y. Alternativamente, possiamo dire che l'$y\%$ degli elementi ha un valore inferiore a x.

Osservando il grafico sopra, possiamo dire che:

- Circa il 40% dei soggetti ha un peso inferiore alle ≈ 33 libre (circa $66kg$);
- Circa l'80% dei soggetti ha un peso inferiore ≈ 42 libre (circa $84Kg$);

Le ECDF tornano utili per verificare graficamente se due fenomeni hanno distribuzioni simili. Possiamo ad esempio usarle per confrontare le distribuzioni dei pesi di uomini e donne:



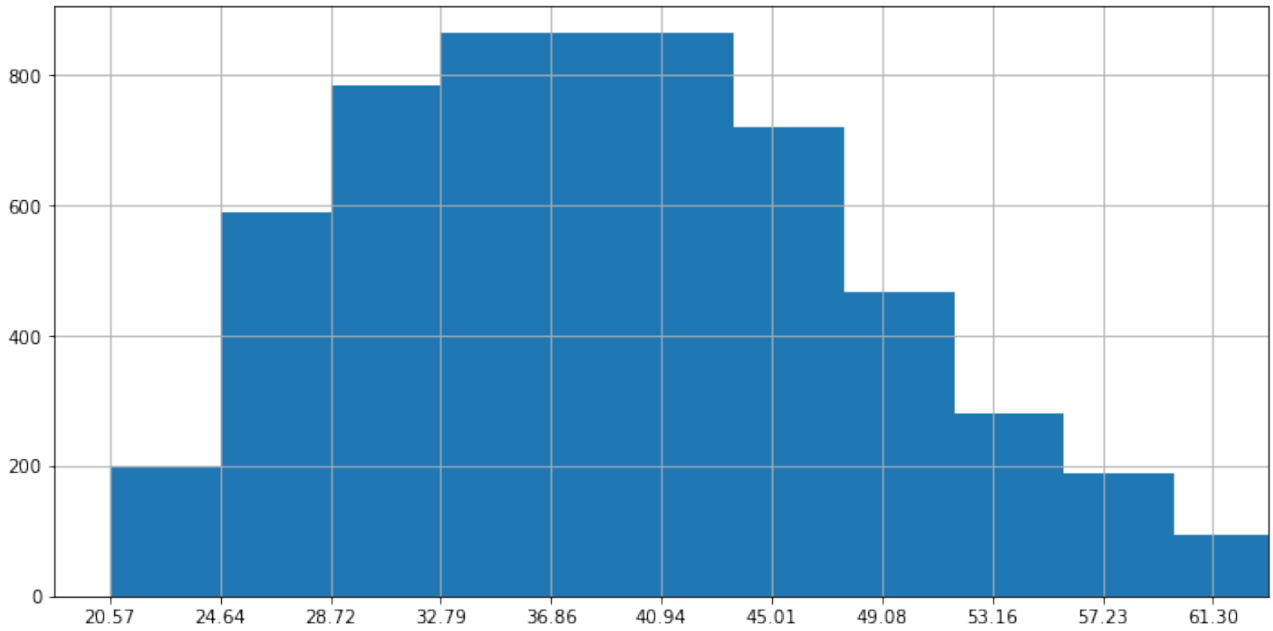
Osservando il grafico sopra, possiamo dire che:

- Circa il 40% degli uomini ha un peso inferiore agli $80Kg$;
- Circa il 40% delle donne ha un peso inferiore ai $60Kg$;
- Il 100% degli uomini ha un peso inferiore ai $140Kg$;
- Il 100% delle donne ha un peso inferiore ai $130Kg$.

In generale, il grafico sopra ci dice che gli uomini tendono a essere più pesanti delle donne.

10.3. Istogrammi

Abbiamo visto che i diagrammi a barre delle frequenze diventano poco chiari quando le variabili sono continue (es. nel caso dei pesi). In questi casi, per ridurre l'influenza del rumore, è possibile utilizzare gli istogrammi. Un istogramma divide il range dei dati in un certo numero di "bin" e riporta per ogni bin il numero di valori che ricadono in quell'intervallo. Di seguito l'istogramma dei pesi nel dataset pesi-altezze:



Ogni "bin" dell'istogramma copre un determinato range. Pratica comune è quella di suddividere il range dei dati, dal minimo al massimo, in un determinato numero di bin della stessa larghezza.

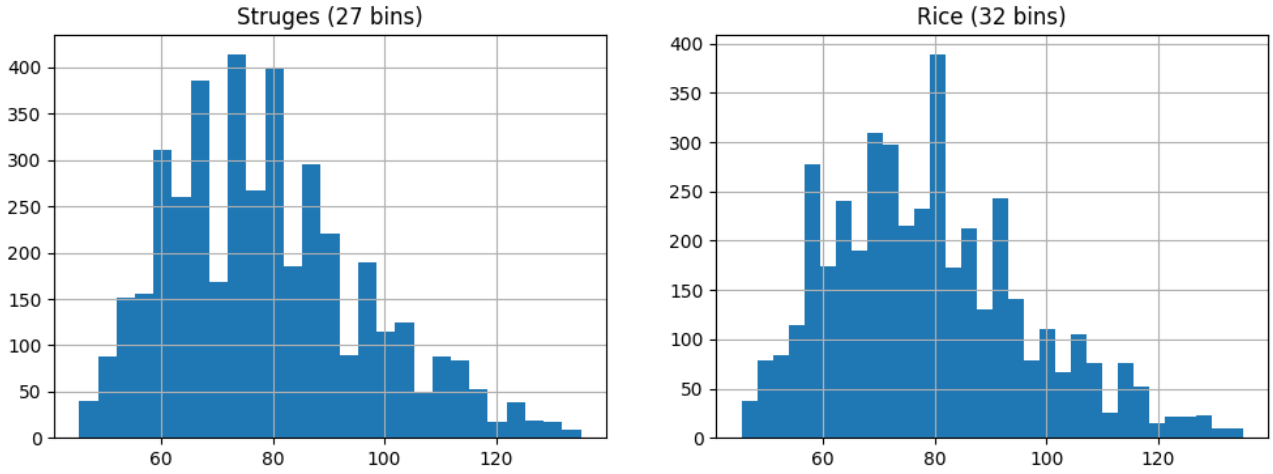
10.3.1. Scegliere il numero di bin: Struges e Rice

Il numero di bin può essere specificato arbitrariamente o determinato a seconda del risultato grafico che si vuole ottenere. Esistono due criteri euristici per trovare dei valori di partenza:

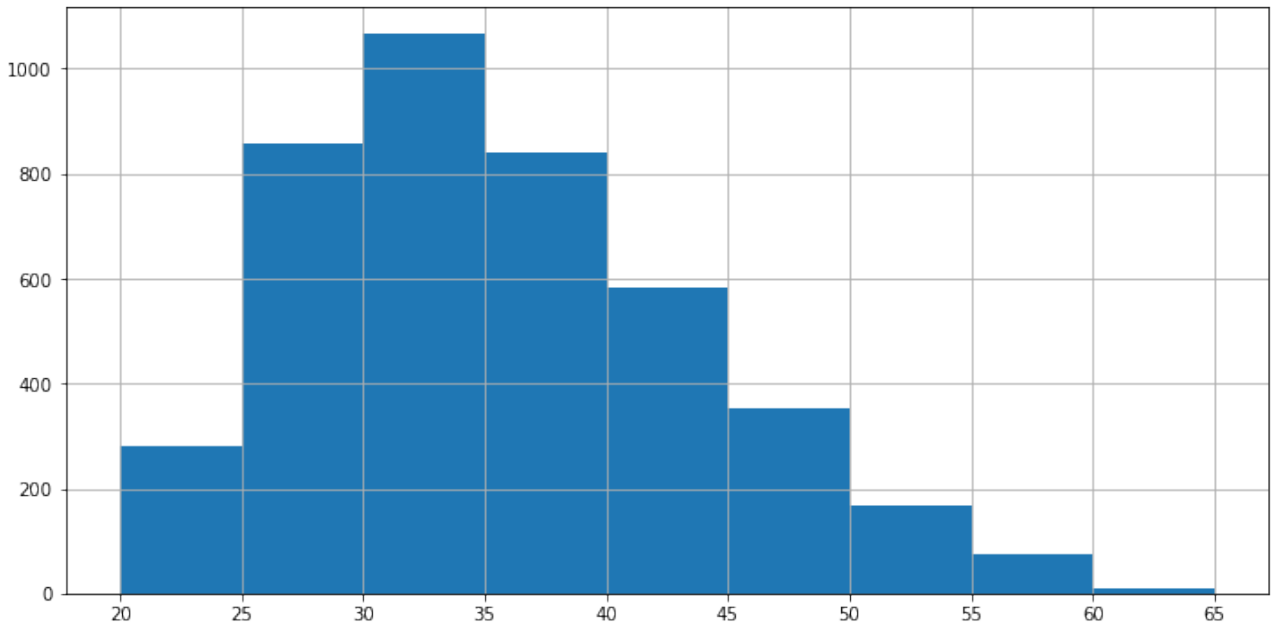
- Struges: $\#bins = 3.3 \log(n)$
- Rice: $\#bins = 2 \cdot n^{1/3}$

Dove n è il numero di elementi.

Va notato che il numero di bin, può cambiare il risultato grafico. Di seguito due esempi ottenuti calcolando il numero di bin con i due criteri considerati:

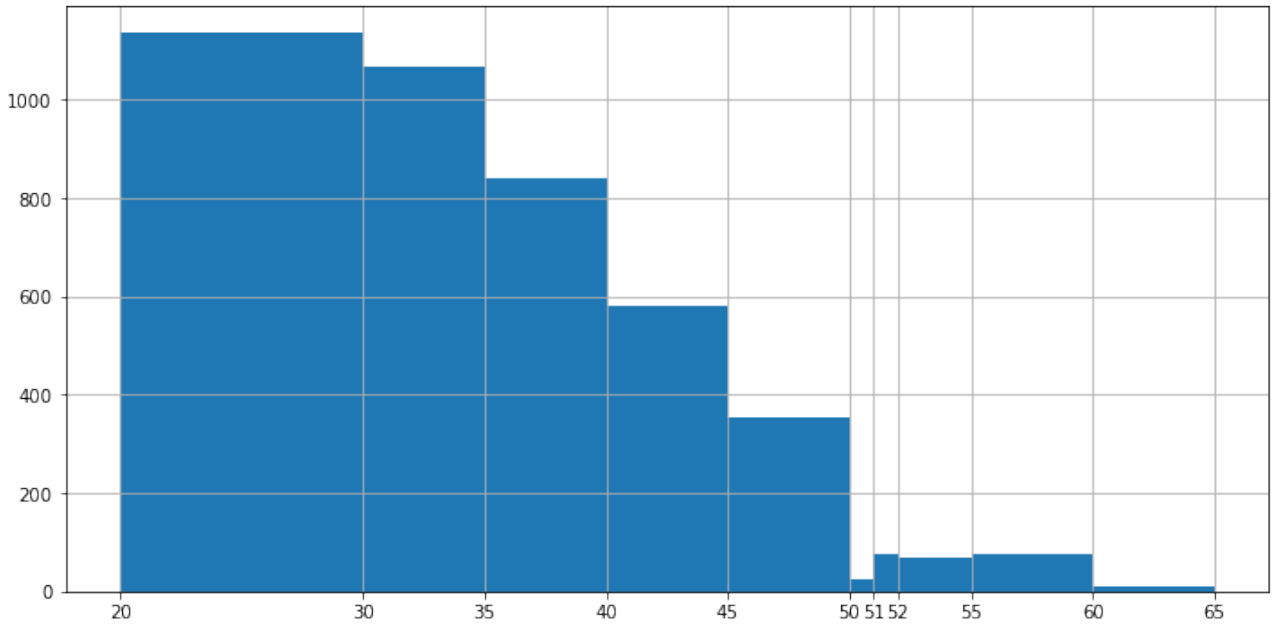


Alternativamente, possiamo definire dei range arbitrari per i nostri bin. Ad esempio, considerando i seguenti limiti per i bin `[20,25,30,35,40,45,50,55,60,65]`, otterremmo il seguente istogramma:



L'istogramma mostrato sopra riporta le frequenze assolute per ogni bin e ci permette di rispondere a domande del genere "quanti soggetti hanno un peso compreso tra 30 e 35 libbre"?

In casi particolari, è possibile anche definire istogrammi con bin di dimensioni variabili. Ad esempio:



10.3.2. Istogrammi di densità

Un istogramma può essere utilizzato anche per approssimare una Probability Density Function. L'istogramma delle frequenze assolute mostrato sopra, tuttavia, non ci permette di ragionare in termini probabilistici.

Ad esempio, non ci permette di dire qual è la probabilità che un soggetto abbia un peso contenuto tra 30 e 40 libbre. Se avessimo la PDF della popolazione dalla quale è stata estratto il campione, potremmo rispondere a questa domanda calcolando l'integrale:

$$\int_{30}^{40} pdf(x)dx$$

Possiamo costruire un **istogramma di densità**, che approssimi in maniera discreta la PDF che cerchiamo. In pratica, vogliamo che l'area sottesa dal bin di "bordi" $[30, 40[$ contenga un valore che approssimi l'integrale della PDF:

$$\int_{30}^{40} pdf(x)dx \approx b_j \cdot w_j$$

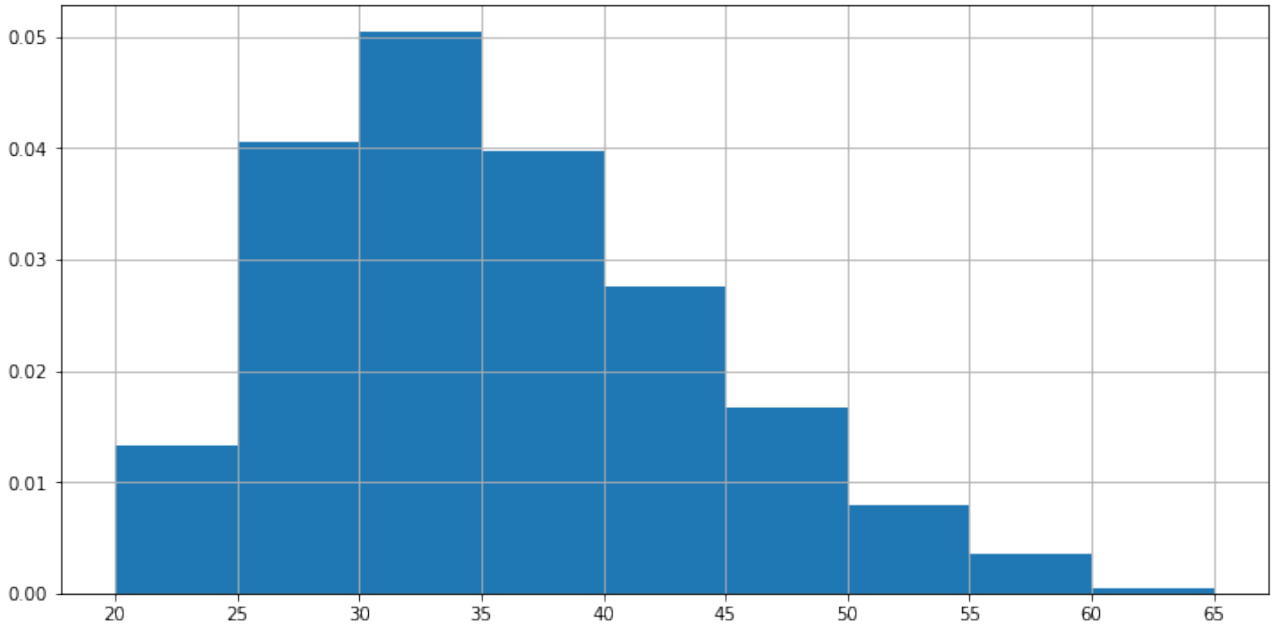
dove j indica il bin di bordi $[30, 40[$, w_j rappresenta la sua larghezza (

$40 - 30 = 10$) e b_j rappresenta la sua altezza (il valore del bin). Sotto queste condizioni, vale dunque la seguente proprietà:

$$\sum_{i=0}^{n} b_i \cdot w_i = \int pdf(x)dx = 1$$

dove n è il numero totale di bin.

Il risultato è graficamente molto simile a quello ottenuto in precedenza, ma cambia la scala sull'asse delle y:



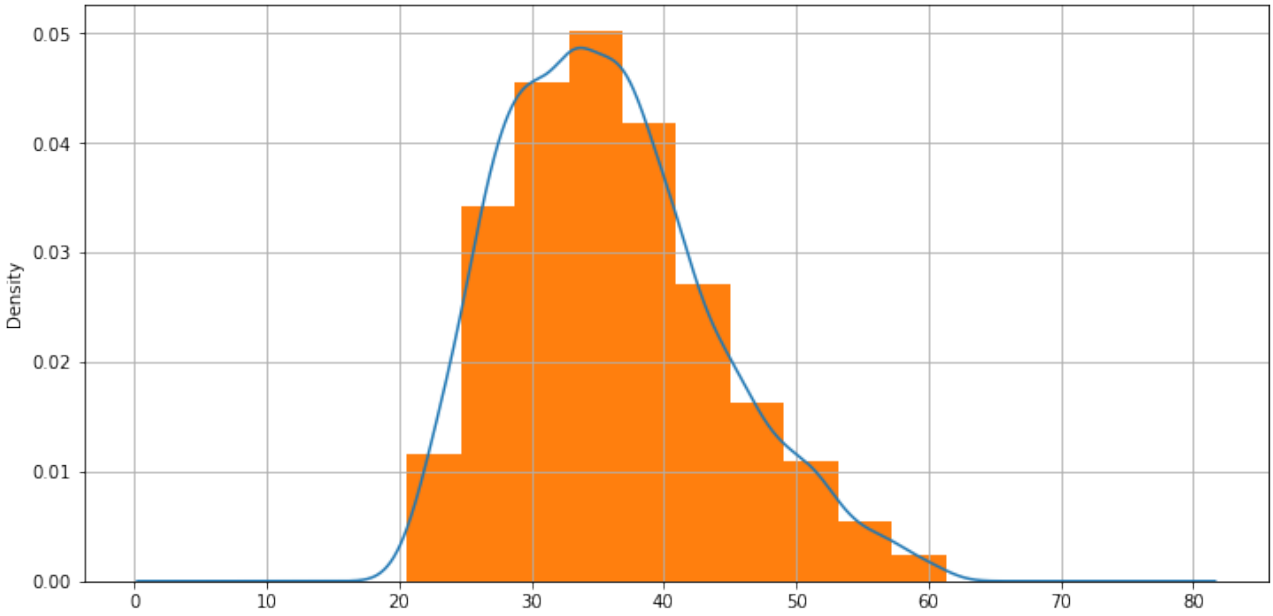
10.3.3. Stima della densità

Uno svantaggio degli istogrammi è che categorizzano dei dati continui in maniera arbitraria mediante dei bin. La scelta degli intervalli dei bin cambia l'aspetto finale dell'istogramma.

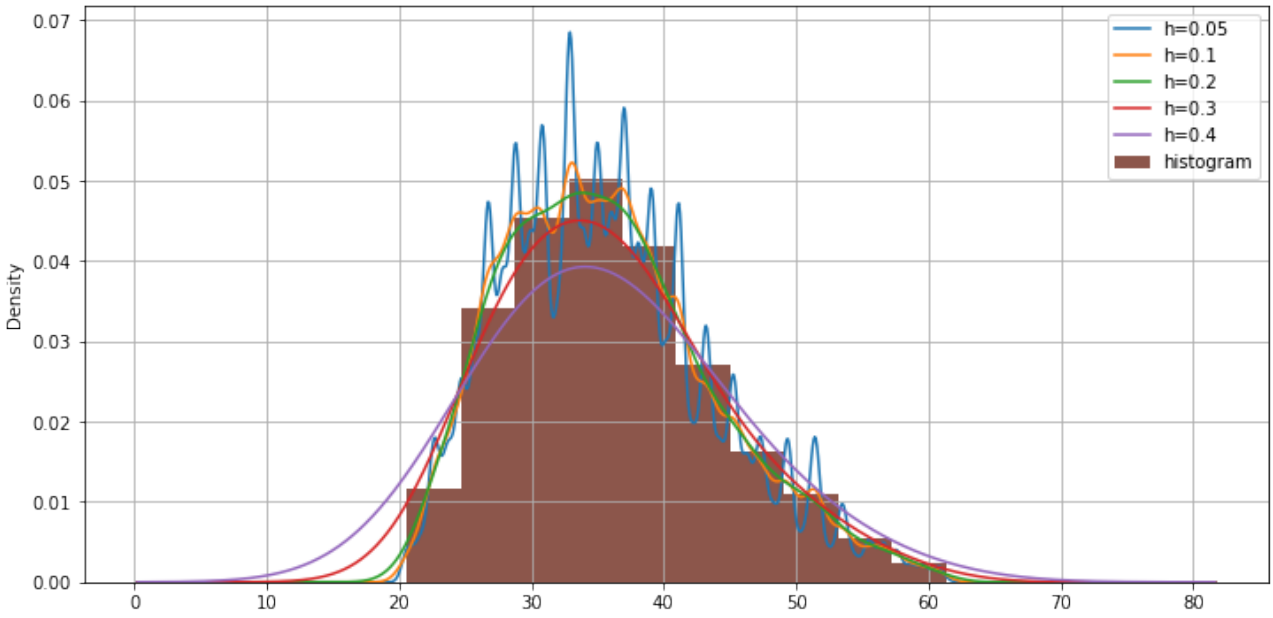
La stima della densità cerca di risolvere questo problema ottenendo una versione "continua" dell'istogramma. Invece di suddividere l'asse delle x in bin, la stima della densità calcola un valore per ciascun punto dell'asse delle x, ottenendo così una rappresentazione continua.

Vedremo meglio come si effettua la stima di densità più in là nel corso. Nel caso uni-dimensionale, può essere calcolata come mostrato nell'approfondimento di seguito (opzionale). Per adesso, sappiamo che si può calcolare la stima della densità con una funzione apposita (lo vedremo a laboratorio) che dipende da un unico parametro di **bandwith** che determina la "sensibilità ai dettagli" della stima della densità.

Confrontiamo la stima di densità dei pesi nel nostro dataset di pesi-altezze con il relativo istogramma:



Se nel caso degli istogrammi cambiare il numero di bin cambiava il risultato grafico, qui è cambiare la bandwidth a cambiare il risultato grafico. Il grafico seguente mostra diversi esempi di stima di densità con diversi valori di bandwidth:



10.3.3.1. Calcolo della stima di densità con Kernel di Epanechnikov (Opzionale)

Il calcolo avviene mediante la seguente formula:

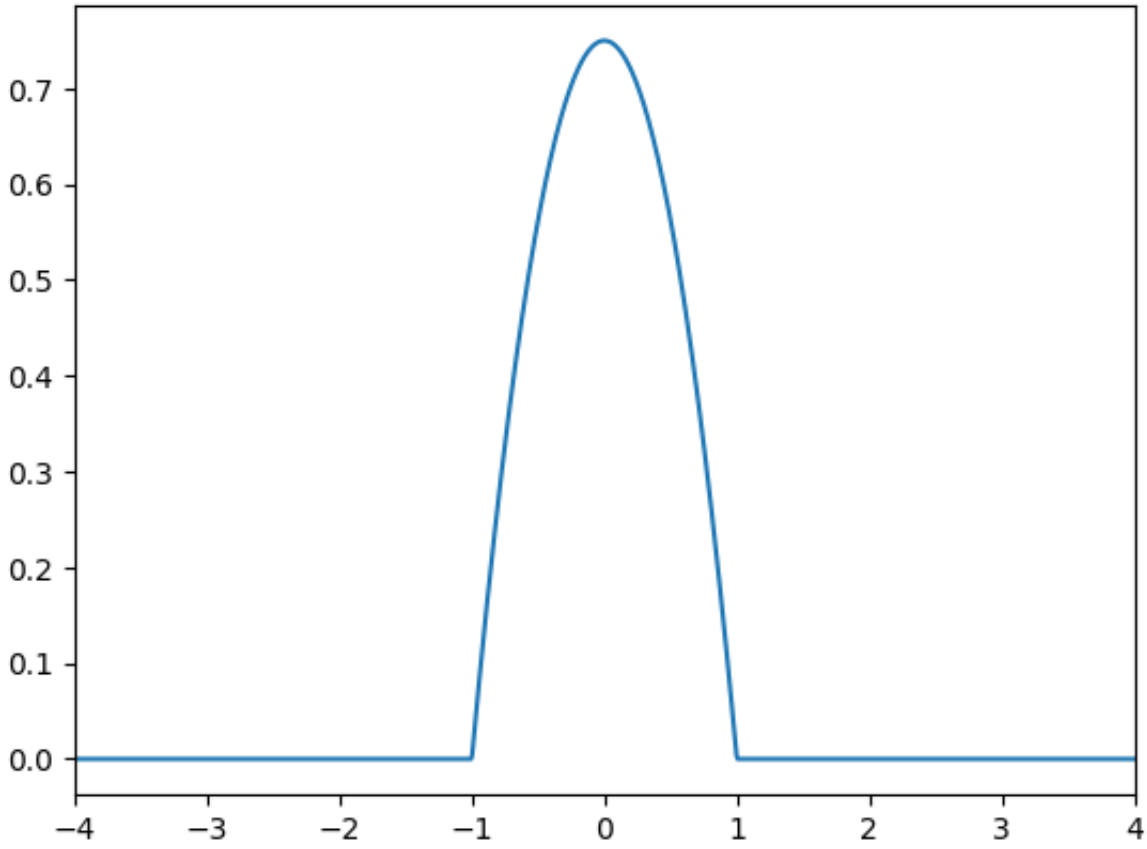
$$\hat{f}_n(x) = \frac{1}{nh} \sum_{i=1}^{n} K(\frac{x - x_i}{h}), h > 0$$

Dove n è la dimensione del campione, h è un parametro detto "bandwidth" e K è una funzione "kernel" che determina quanto gli elementi del campione devono contribuire alla stima nel punto x, dipendentemente dalla loro distanza da x. Una scelta comune di kernel è quello di Epanechnikov:

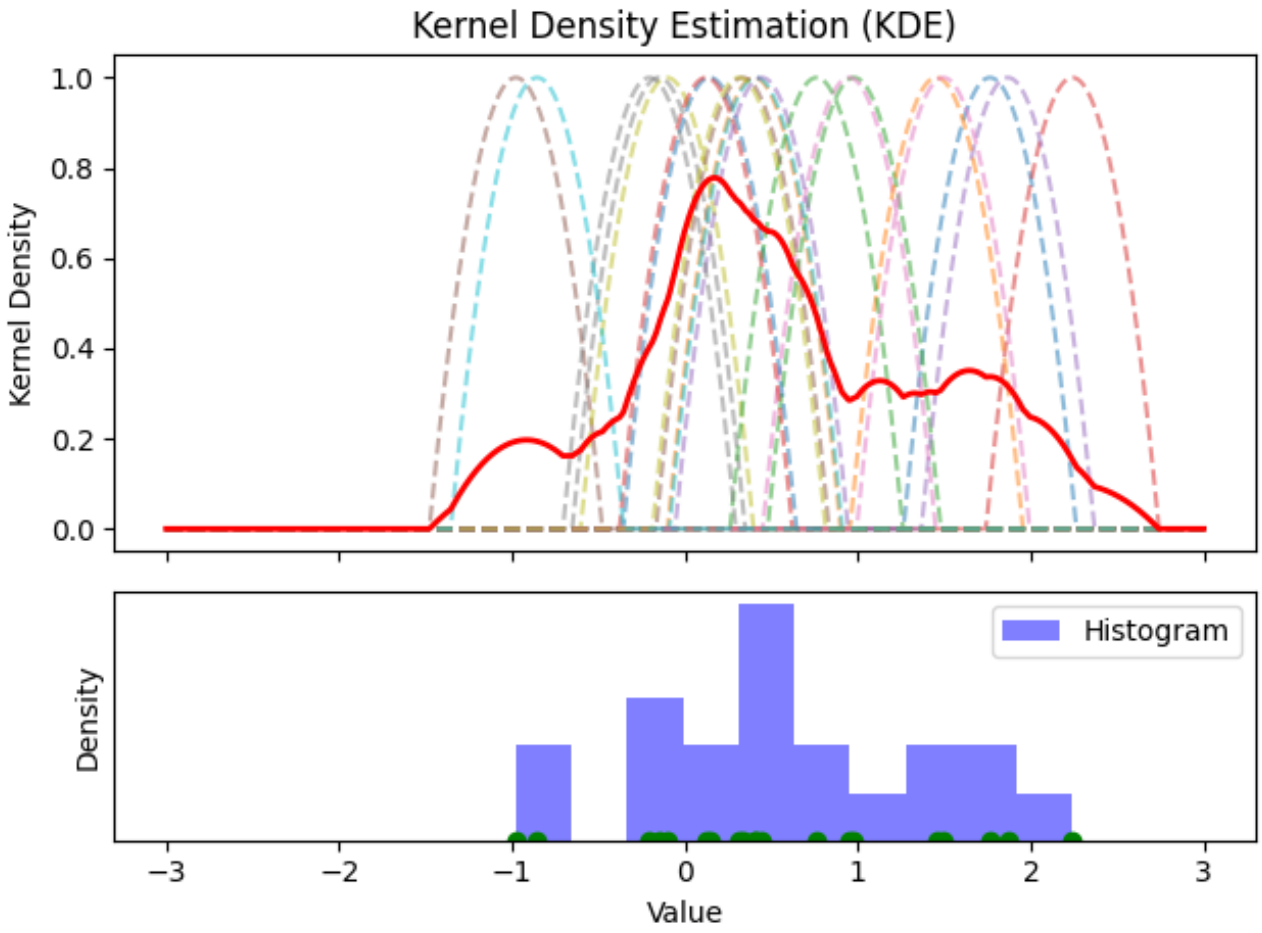
$$K(x) = \begin{cases} \frac{3}{4}(1 - x^2) & \text{if } |x| \leq 1 \\ 0 & \text{otherwise.} \end{cases}$$

In pratica il Kernel di Epanechnikov ha la seguente forma:

```
(-4.0, 4.0)
```

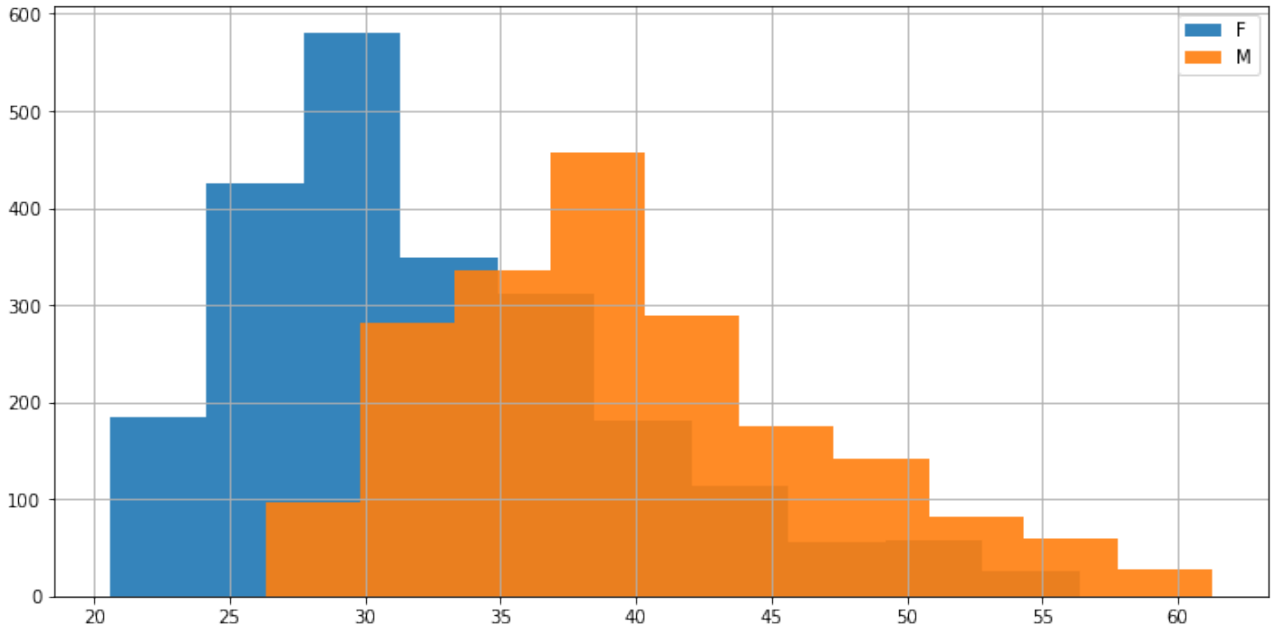


Facendo “scorrere” questo kernel su diversi punti dell’asse delle x e sommando i contributi, si ottiene la stima di densità finale:



10.3.4. Confrontare campioni mediante istogrammi

Gli istogrammi possono essere utili per comparare campioni. In grafico che segue confronta i pesi di uomini e donne nel nostro campione pesi-altezze:



10.4. Referenze

- Capitolo 2 di: Heumann, Christian, and Michael Schomaker Shalabh. Introduction to statistics and data analysis. Springer International Publishing Switzerland, 2016.

Previous
9. Introduzione a Pandas

Next
11. Misure di Tendenza Centrale, Dispersione e Forma