Statistica I

Unità H: simmetria, curtosi & multimodalità

Tommaso Rigon

Università Milano-Bicocca

Unità H

Argomenti affrontati

- Concetto di simmetria
- Indici di asimmetria di Pearson e Bowley
- Concetto di curtosi
- Indice di curtosi di Pearson
- Cenni alla multimodalità

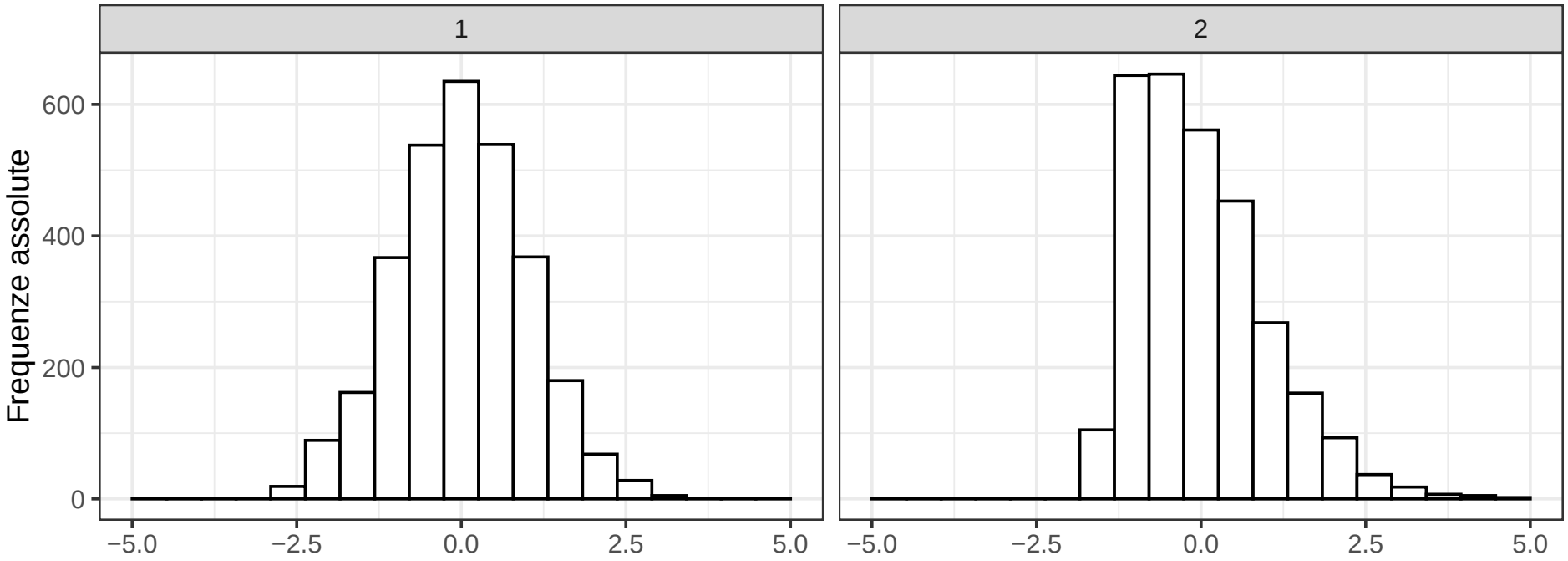
Riferimenti al libro di testo

- §6.1 — §6.3
- **Nota**. Nel libro di testo sono presenti vari altri indici di asimmetria.

La simmetria

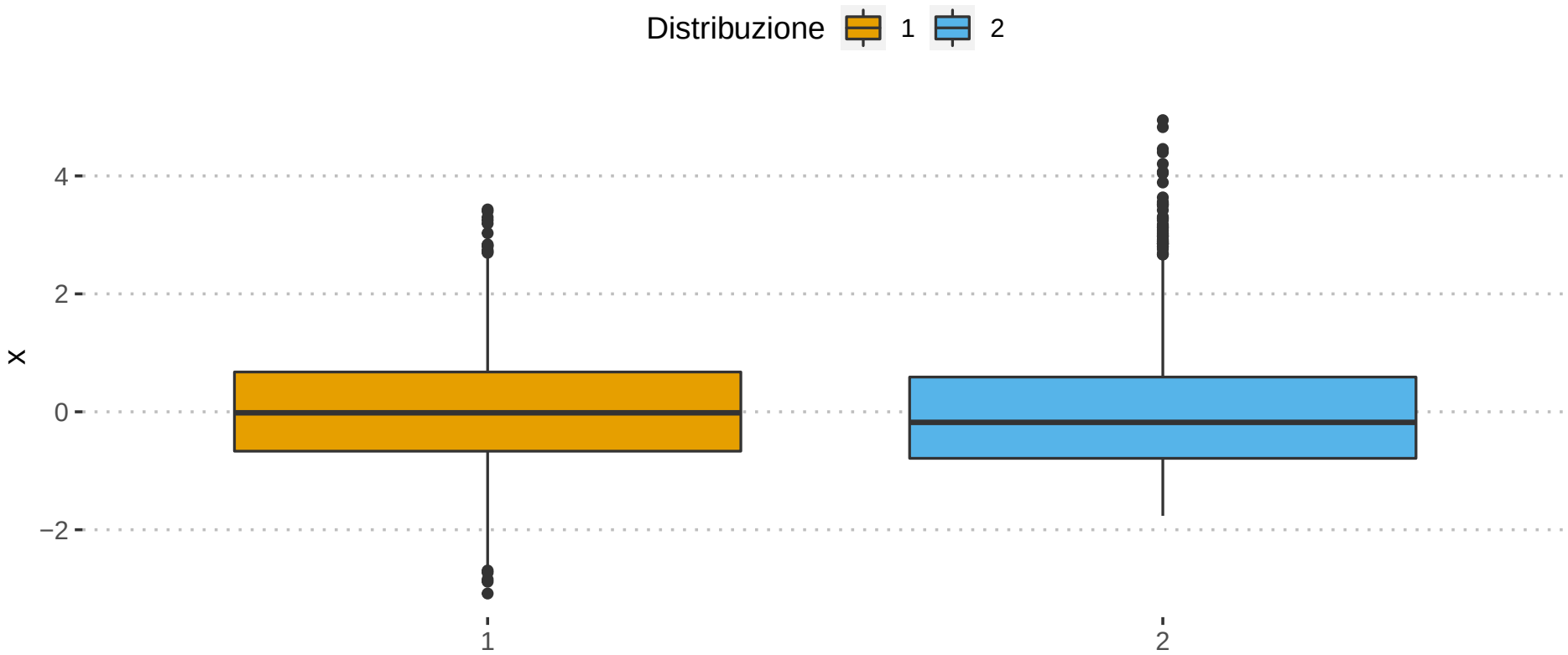
- Nelle slide che seguono consideriamo due insiemi di **dati standardizzati**, ovvero ottenuto come descritto alla fine dell'unità F.
- Per definizione, questi insiemi di dati hanno media pari a zero e varianza pari a 1.
- I due insiemi di dati sono perciò abbastanza omogenei per quanto riguarda posizione e variabilità.
- Nonostante media e varianza siano uguali, le due distribuzioni sono evidentemente molto diverse.

Due insiemi di dati standardizzati



- La prima distribuzione è sostanzialmente simmetrica rispetto allo zero.
- Nel secondo caso, la **coda** verso i valori alti è molto più pronunciata della coda verso i valori bassi. Questa distribuzione viene detta **asimmetrica positiva**. Nel caso opposto (coda sinistra maggiormente pronunciata) verrebbe detta **asimmetrica negativa**.

Due insiemi di dati standardizzati



Indici di asimmetria

- La simmetria è definita qualitativamente come la **specularità** della distribuzione rispetto ad un asse.
- Vogliamo quindi quantificare l'assenza di simmetria, ovverso l'asimmetria, tramite degli indici.
- Un primo e semplice indice di asimmetria potrebbe basarsi sul **confronto tra media e mediana**. Infatti se una distribuzione è simmetrica, allora (media) = (mediana).
- Sulla base di questo indicatore, definiamo una distribuzione asimmetrica positiva se $\bar{x} - \text{Me} > 0$ e asimmetrica negativa se $\bar{x} - \text{Me} < 0$.
- **Nota**. Esistono distribuzioni non simmetriche tali che (media) = (mediana). Si veda **Esempio 6.3** (pag. 163) del libro di testo.

Indice di asimmetria di Pearson

- La misura di asimmetria di uso più comune è il cosiddetto indice di asimmetria standardizzato di Pearson.

Indice di asimmetria di Pearson. L'indice di asimmetria dei dati $x_1, \ldots, x_n$ è

$$\gamma = \frac{1}{\text{sqm}(x)^3} \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^3 = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{\sigma} \right)^3 .$$

- Se i dati si distribuiscono in maniera **simmetrica intorno alla media** i termini positivi e negativi nella sommatoria si compenseranno tra di loro e quindi avremo $\gamma = 0$.

- Viceversa, sulla base di questo indicatore, definiamo una distribuzione asimmetrica positiva se $\gamma > 0$ e asimmetrica negativa se $\gamma < 0$.

- Infatti, nei casi di asimmetria positiva i termini positivi predomineranno e quindi l'indice assumerà valori positivi. Opposta la situazione nei casi di asimmetria negativa.

Proprietà indice di asimmetria di Pearson

- L'indice di asimmetria Pearson è **standardizzato**. Si noti infatti che γ si può calcolare come segue

$$\gamma = \frac{1}{n} \sum_{i=1}^{n} z_i^3,$$

dove $z_1, \ldots, z_n$ rappresentano i dati standardizzati.

- L'indice, per costruzione, è invariante rispetto a trasformazioni lineari dei dati.

- In altri termini, otteniamo lo stesso risultato sia lavorando con i dati originali che considerando la trasformazione lineare $y_i = a + bx_i$, per $i = 1, \ldots, n$.

- **Esercizio**. Si verifichi questa proprietà.

Indice di asimmetria di Bowley

- Una misura di asimmetria alternativa, attribuita a A.L. Bowley e a G.U. Yule, si basa sui quartili.

- **Indice di asimmetria di Bowley**. L'indice di asimmetria dei dati $x_1, \ldots, x_n$ è

$$B = \frac{(\mathcal{Q}_{0.75} - \text{Me}) + (\mathcal{Q}_{0.25} - \text{Me})}{\mathcal{Q}_{0.75} - \mathcal{Q}_{0.25}} = \frac{\mathcal{Q}_{0.75} - 2\text{Me} + \mathcal{Q}_{0.25}}{\mathcal{Q}_{0.75} - \mathcal{Q}_{0.25}}.$$

- Nei casi in cui i dati si distribuiscano in maniera **simmetrica intorno alla mediana** i termini a numeratore si compenseranno tra di loro e quindi avremo $B = 0$.

- Viceversa, sulla base di questo indicatore, definiamo una distribuzione asimmetrica positiva se $B > 0$ e asimmetrica negativa se $B < 0$.

- Nel caso di asimmetria positiva, la differenza tra $\mathcal{Q}_{0.75}$ e Me sarà maggiore alla differenza tra Me e $\mathcal{Q}_{0.25}$. Opposta la situazione nei casi di asimmetria negativa.

- L'indice di Bowley assume valore minimo in -1 quando il terzo quartile coincide con la mediana e valore massimo in 1 quando il primo quartile coicide con la mediana.

Due insiemi di dati standardizzati

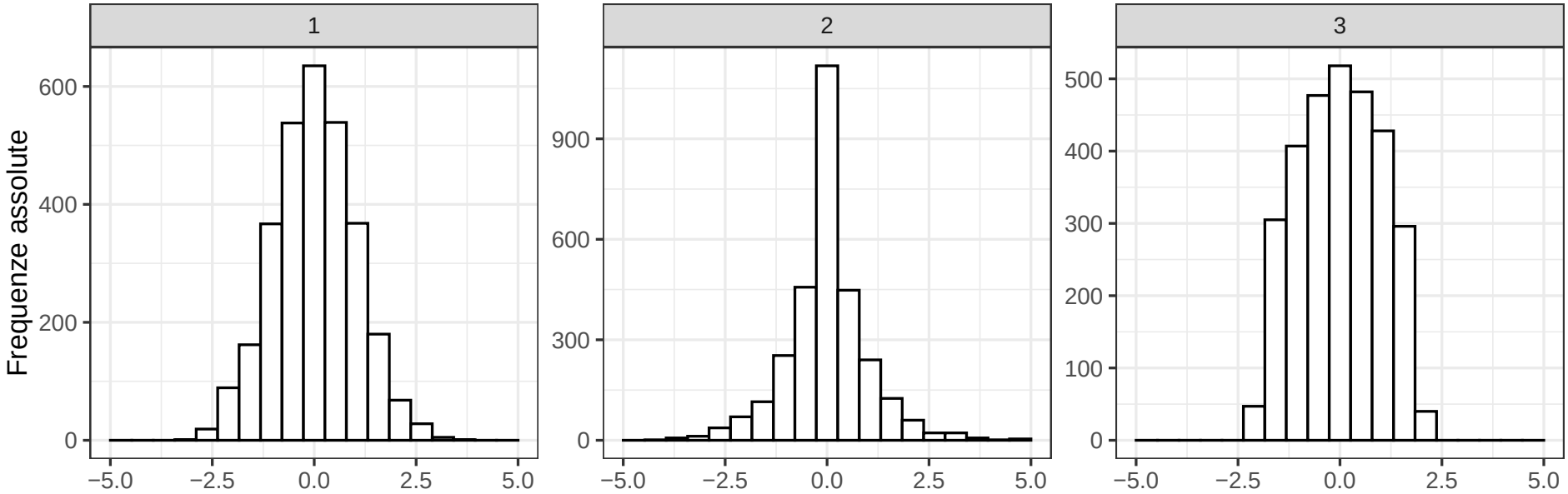
	Insieme di dati 1	Insieme di dati 2
Media	0	0
Varianza	1	1
Media - Mediana	0.017	0.179
Asimmetria di Pearson γ	0.034	0.949
Asimmetria di Bowley B	0.030	0.115

- L'insieme di dati 1 è essenzialmente **simmetrico**: tutti gli indici sono circa pari a zero.
- Tutti gli indici suggeriscono la presenza di **asimmetria positiva** nell'insieme di dati 2, come del resto si poteva evincere dall'istogramma.

La curtosi

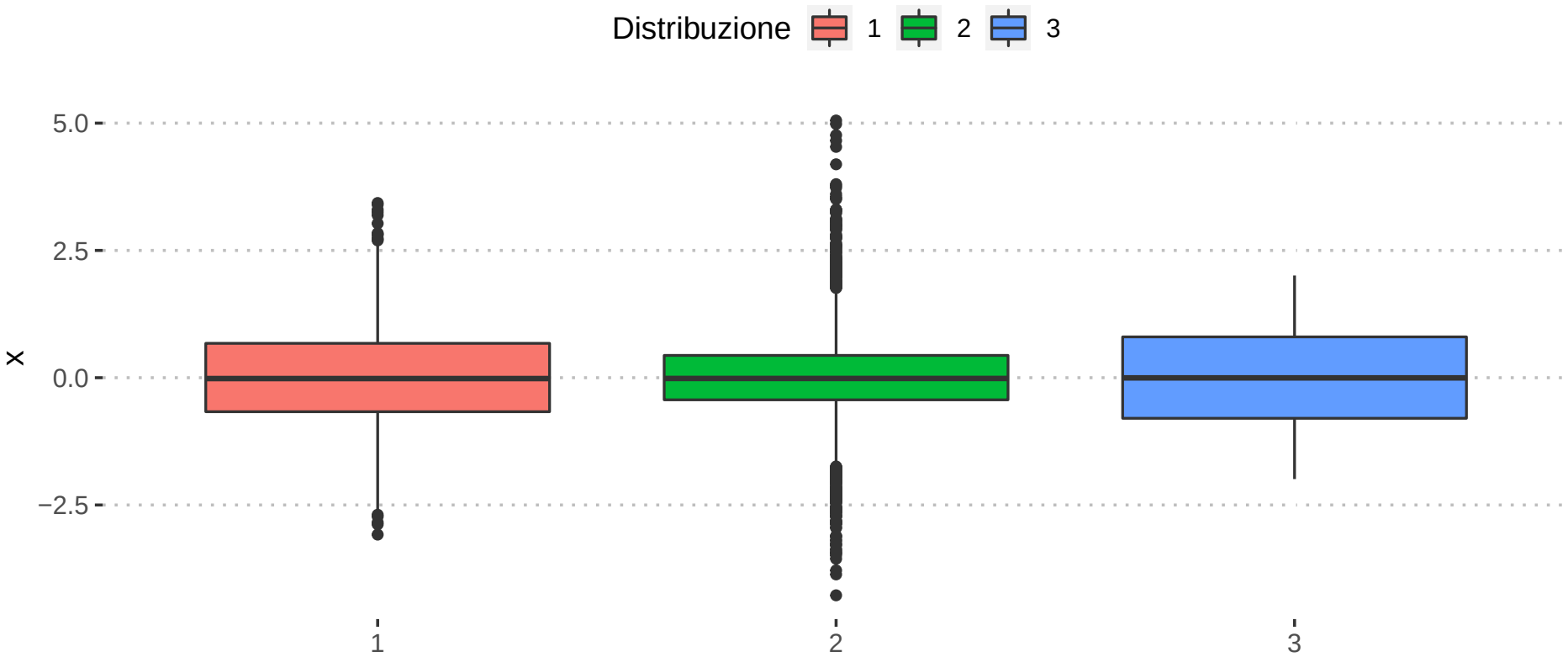
- Nei grafici nelle seguenti confrontiamo tre insiemi di dati standardizzati.
- Le tre distribuzioni sono sostanzialmente simmetriche, come si evince dagli istogrammi e dagli indici di asimmetria.
- Nonostante l'uguaglianza delle medie, delle varianze e la simmetria, queste tre distribuzioni sono **molto diverse**.
- Queste distribuzioni differiscono per un quarto aspetto, che chiameremo **curtosi**.

Tre insiemi di dati standardizzati



- La seconda distribuzione ha delle code più "pesanti" ed è più appuntita della prima.
- Viceversa, la terza distribuzione ha le code più leggere ed è meno appuntita della prima.
- Questa caratteristica, ovvero il maggiore o minore peso delle code e maggiore o minore "appuntimento" (a parità di variabilità), è spesso indicata con il termine **curtosi**.

Tre insiemi di dati standardizzati



Indice di curtosi di Pearson

- La misura di curtosi di uso più comune è il cosiddetto indice di curtosi **standardizzato** di Pearson.
- **Indice di curtosi di Pearson**. L'indice di curtosi dei dati $x_1, \ldots, x_n$ è

$$\kappa = \frac{1}{\text{sqm}(x)^4} \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^4 = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{\sigma} \right)^4 .$$

- L'indice di curtosi è tale che $\kappa \geq 0$ ed è pari a zero solamente se i dati sono costanti.
- Si osservi che κ essere visto come un rapporto tra due indici di variabilità.
- L'indice a numeratore è scelto in maniera tale da essere più sensibile alla presenza di code pesanti dell'indice al denominatore.

Proprietà indice di curtosi di Pearson

- L'indice di curtosi Pearson è **standardizzato**. Si noti infatti che κ si può calcolare come segue

$$\kappa = \frac{1}{n}\sum_{i=1}^{n} z_i^4,$$

dove $z_1, \ldots, z_n$ rappresentano i dati standardizzati.

- Pertanto l'indice è invariante rispetto a trasformazioni lineari dei dati, come nel caso dell'indice di asimmetria γ.

- Per ragioni legate al calcolo delle probabilità, il valore

$$\kappa = 3$$

viene convenzionalmente preso come riferimento.

- Di conseguenza quando $\kappa > 3$ si parla, ad esempio, di **eccesso di curtosi**.

Tre insiemi di dati standardizzati

	Insieme di dati 1	Insieme di dati 2	Insieme di dati 3
Media	0	0	0
Varianza	1	1	1
Asimmetria γ	0.034	0.162	-0.017
Curtosi di Pearson κ	2.961	5.520	2.011

- Tutti e tre gli insieme di dati hanno la stessa media e varianza. Inoltre, sono sostanzialmente simmetriche.
- La prima distribuzione ha curtosi molto vicino a 3. Questa forma viene presa convenzionalmente come riferimento.
- Viceversa, le distribuzioni 2 e 3 sono, rispettivamente, più o meno appuntite. Questo aspetto viene registrato dall'indice κ.

Old Faithful Geyser di Yellowstone

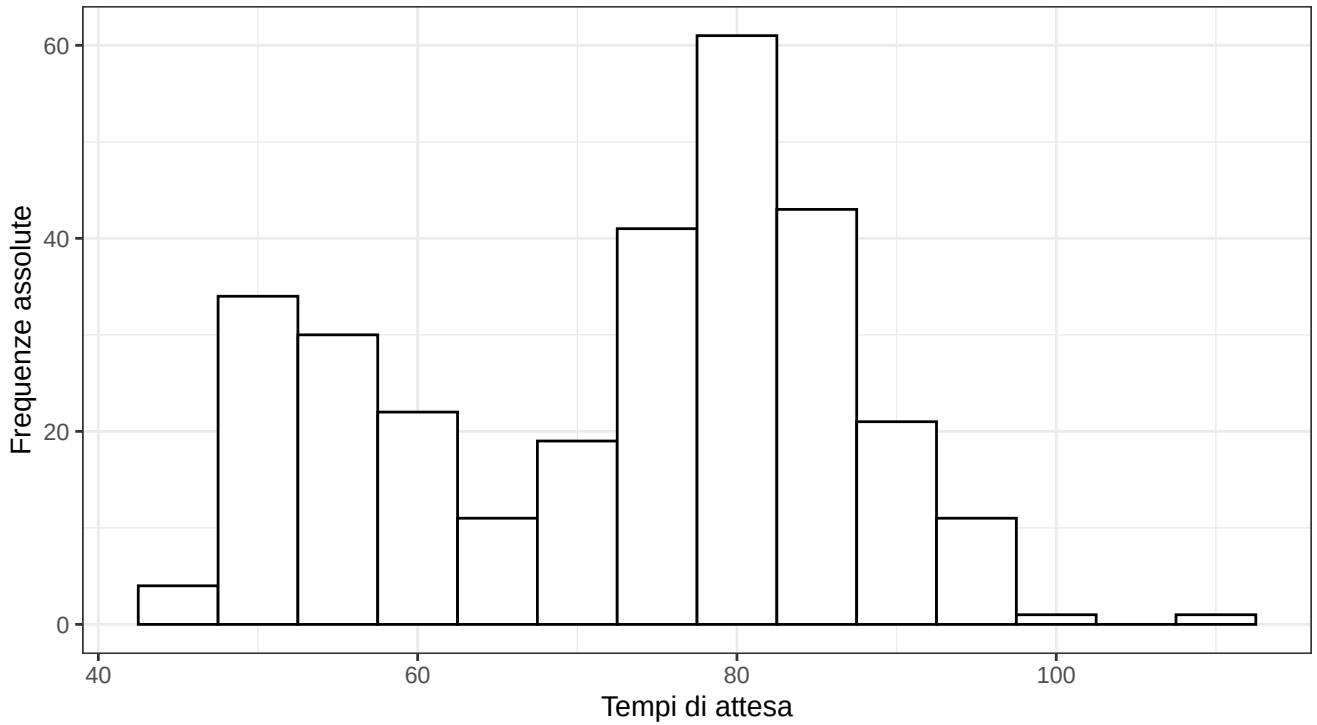
- L'**Old Faithful Geyser** si trova nel parco nazionale di Yellowstone, Wyoming, U.S.A. ed erutta ad intervalli regolari.
- Siamo interessati descrivere la distribuzione dei **tempi di attesa** tra un'eruzione e quella successiva, per poter fornire indicazioni a turisti in visita.
- Le $n = 299$ osservazioni a nostra disposizione sono state raccolte tra il 1 ed il 15 Agosto del 1985.

Old Faithful Geyser: statistiche descrittive

	Tempo di attesa (minuti)
Minimo	43
Primo quartile	59
Media	72.31
Mediana	76
Terzo quartile	83
Massimo	108

- Queste statistiche descrittive **sembrano** suggerire che il tempo di attesa tra un'eruzione e quella successiva sia mediamente 72 minuti.
- Inoltre, la maggior parte delle attese **sembrano** durare tra i 59 e gli 83 minuti.
- Tuttavia, l'ispezione dell'istogramma rivela una storia molto diversa.

Old Faithful Geyser: istogramma



- La forma della distribuzione dei tempi di attesa presenta **due picchi**: uno più basso intorno ai 50 minuti ed un secondo più alto intorno agli 80.
- Gli indici di posizione considerati non riescono a descrivere questo comportamento.
- La media identifica il centro della distribuzione in un punto dove ci sono **pochi dati**.

Indici di posizione e multimodalità

- La distribuzione dei tempi di attesa dell'Old Faithful geyser è un esempio di **distribuzione bimodale**.
- Quando la distribuzione presenta un unico "picco" si dice, invece, unimodale.
- In presenza di distribuzioni bimodali o multimodali (più di un picco), bisogna fare molta **attenzione**: gli indici di posizione potrebbero non essere particolarmente rilevanti.
- In questo caso, la media e la mediana **non** sono particolarmente **interessanti**. Ben più utile sarebbe invece capire la posizione del primo e del secondo picco.
- Identificare la precisa posizione dei picchi non è un problema semplice e richiede strumenti statistici leggermente più avanzati, che non vedremo in questo corso.