# Statistica I

## Esercitazione 3: variabilità, istogrammi, boxplot, simmetria, curtosi

**Tommaso Rigon**

**Università Milano-Bicocca**

# La saggezza della folla è affidabile?

- Nel 1907 lo scienziato ed inventore **Francis Galton**, cugino di Charles Darwin, scrisse una lettera alla prestigiosa rivista scientifica **Nature**, intitolata "*Vox Populi*".
- Francis Galton era stato ad una mostra di bestiame. Era stato indetto un **concorso** per indovinare il peso della carne, dopo la macellazione, di un **grande bue**.
- La partecipazione costava 6 penny e Galton riuscì a procurarsi ben $n = 787$ dei biglietti acquistati. Calcolò la **media** delle varie stime, ovvero 547 kg.
- Il valore medio era notevolmente vicino al **peso reale** di 543 kg, sebbene la maggior parte dei concorrenti avesse fornito una stima molto meno precisa.
- Questo metodo per prendere decisioni è spesso chiamato "**saggezza della folla**".
- Siamo interessati a verificare **se** il fenomeno della saggezza della folla sia replicabile, tramite un esperimento.

# Galton, F. (1907), Vox populi, *Nature*, 1949(75)

## VOX POPULI.

IN these democratic days, any investigation into the trustworthiness and peculiarities of popular judgments is of interest. The material about to be discussed refers to a small matter, but is much to the point.

A weight-judging competition was carried on at the annual show of the West of England Fat Stock and Poultry Exhibition recently held at Plymouth. A fat ox having been selected, competitors bought stamped and numbered cards, for 6*d*. each, on which to inscribe their respective names, addresses, and estimates of what the ox would weigh after it had been slaughtered and "dressed." Those who guessed most successfully received prizes. About 800 tickets were issued, which were kindly lent me for examination after they had fulfilled their immediate purpose. These afforded excellent material. The judgments were unbiassed by passion and uninfluenced by oratory and the like. The sixpenny fee deterred practical joking, and the hope of a prize and the joy of competition prompted each competitor to do his best. The competitors included butchers and farmers, some of whom were highly expert in judging the weight of cattle; others were probably guided by such information as they might pick up, and by their own fancies. The average competitor was probably as well fitted for making a just estimate of the dressed weight of the ox, as an average voter is of judging the merits of most political issues on which he votes, and the variety among the voters to judge justly was probably much the same in either case.

After weeding thirteen cards out of the collection, as being defective or illegible, there remained 787 for discussion. I arrayed them in order of the magnitudes of the estimates, and converted the *cwt.*, *quarters*, and *lbs.* in which they were made, into lbs., under which form they will be treated.





*Distribution of the estimates of the dressed weight of a particular living ox, made by 787 different persons.*

| Degrees of the length of Array 0°—100° | Estimates in lbs. | Centiles: Observed deviates from 1207 lbs. | Centiles: Normal p.e =37 | Excess of Observed over Normal |
|---|---|---|---|---|
| 5 | 1074 | −133 | −90 | +43 |
| 10 | 1109 | − 98 | −70 | +28 |
| 15 | 1126 | − 81 | −57 | +24 |
| 20 | 1148 | − 59 | −46 | +13 |
| $q_1$ 25 | 1162 | − 45 | −37 | + 8 |
| 30 | 1174 | − 33 | −29 | + 4 |
| 35 | 1181 | − 26 | −21 | + 5 |
| 40 | 1188 | − 19 | −14 | + 5 |
| 45 | 1197 | − 10 | − 7 | + 3 |
| *m* 50 | 1207 | 0 | 0 | 0 |
| 55 | 1214 | + 7 | + 7 | 0 |
| 60 | 1219 | + 12 | +14 | − 2 |
| 65 | 1225 | + 18 | +21 | − 3 |
| 70 | 1230 | + 23 | +29 | − 6 |
| $q_3$ 75 | 1236 | + 29 | +37 | − 8 |
| 80 | 1243 | + 36 | +46 | −10 |
| 85 | 1254 | + 47 | +57 | −10 |
| 90 | 1267 | + 52 | +70 | −18 |
| 95 | 1293 | + 86 | +90 | − 4 |

$q_1$, $q_3$, the first and third quartiles, stand at 25° and 75° respectively.
*m*, the median or middlemost value, stands at 50°.
The dressed weight proved to be 1198 lbs.

- **Articolo**: `https://galton.org/essays/1900-1911/galton-1907-vox-populi.pdf`

# La bottiglia con le biglie di vetro

- Ho riempito una bottiglia con delle **biglie** e ho chiesto alla classe quante fossero.

# Informazioni aggiuntive

- Ho quindi posto nuovamente la domanda agli studenti, fornendo però le seguenti **informazioni aggiuntive**. Gli studenti potevano **rivedere** la loro stima.

## Informazioni aggiuntive

- La formula per il calcolo del **volume di un cilindro** è:

$$(\text{Volume}) = (\text{Area di base}) \times (\text{Altezza}),$$

  dove l'area di base, ovvero l'**area del cerchio**, è pari a:

$$(\text{Area di base}) = \pi(\text{Diametro}/2)^2.$$

- La bottiglia è stata agitata varie volte durante il riempimento, per ridurre il più possibile gli spazi vuoti tra le biglie.

- La bottiglia contiene approssimativamente **1 litro**. Le biglie di vetro sono tutte uguali tra di loro e hanno diametro di **16 mm**, ovvero **2.144 ml** ciascuna.

- È inoltre **noto** (`https://www.nature.com/articles/nature06981`) che l'impacchettamento casuale di sfere ha una **densità** compresa tra il 55% ed il 64%.

# I dati grezzi

## Primo tentativo ($n = 95$)

```
 [1] 250 210 136 250 450 240 251 210  NA  NA 187  96 135 350 210 400
[17] 260 450 219 175 330 287 291 115 270 275 293 177 204 264 220 142
[33] 305 300 201 426 317 168 145 112 250 370 255  NA 149 227 670 213
[49] 270 213 185 190 167 247 235 167  NA 362 291  NA 345 125 300 280
[65] 250 207 190 163 400 297 160 227 157 250 187 184 247 211 187 133
[81]  NA 315  79 335 260 218 248 225 330 197 250 244  NA 189 295
```

## Secondo tentativo ($n = 95$)

```
 [1]  220  280  280  285  267  330  253  200  350  275  466   NA  380
[14]  260   NA  279  134  270  263  213  300  296  277   NA  180  275
[27]  270  279  249  293  330  148  370  300  210  466  269  256  200
[40]  279  298  220  280  351   NA  215 1096  127  280  153  465  280
[53]  235  247  262  289  276  322  280  268  280  220  250  280  280
[66]  226  230  298  275  278  190  250  202  275  423   NA  338   NA
[79]  256   38  390  293  265  300  281  263   NA  280  298  253  275
[92]  279  263  238  100
```

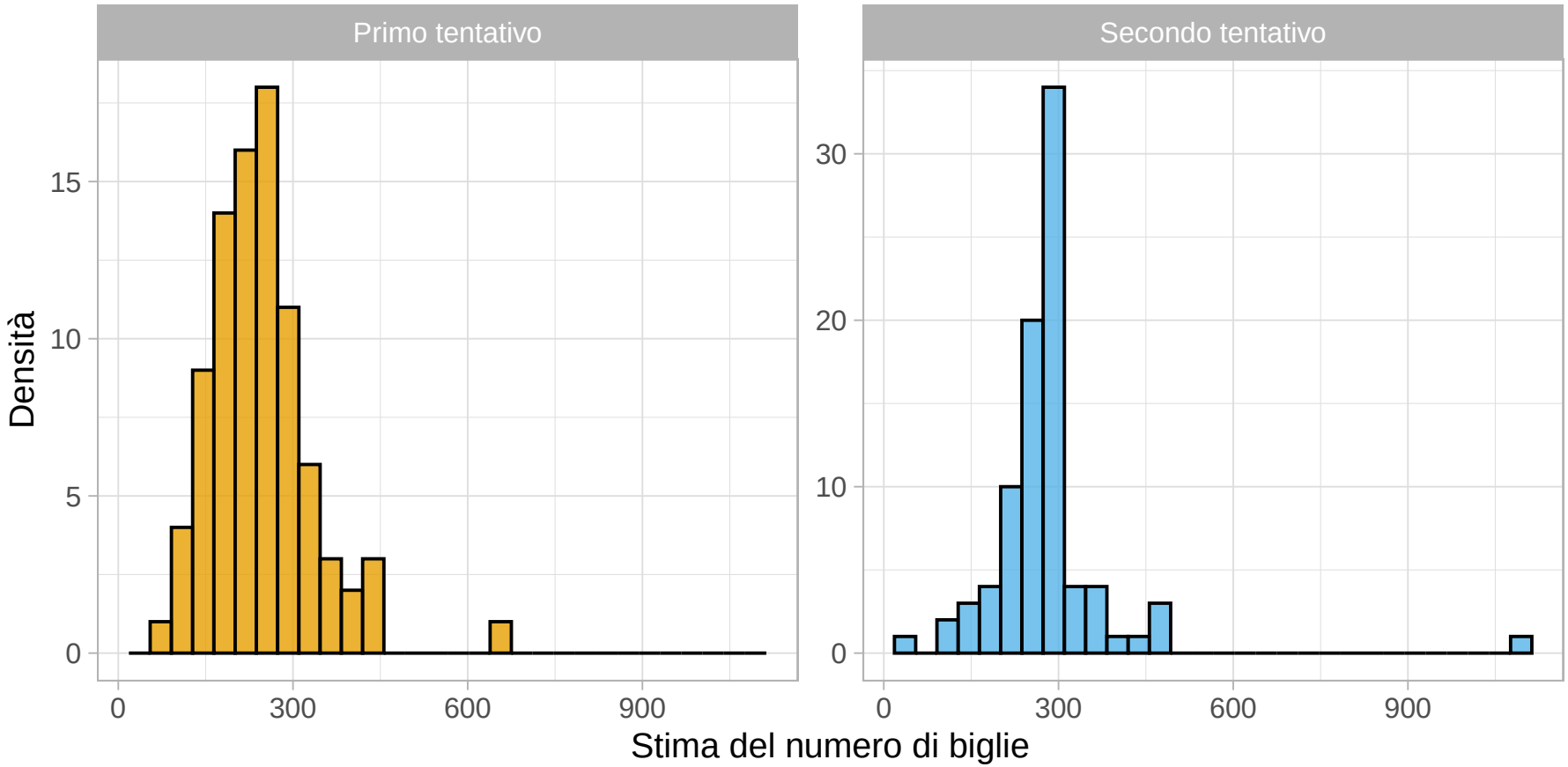
# La pulizia del dataset

- I dati di questa unità presentano delle **complicazioni ulteriori**, tipiche dei **dati reali**.
- Il numero di studenti che si sono connessi per votare sono $n = 95$. Tuttavia, alcune di loro hanno **votato solamente una volta**.
- L'assenza del voto è un **dato mancante** e si indica con la sigla `NA`, dall'inglese "*Not Available*". In questa sede, noi ci limiteremo ad escludere dall'analisi i valori mancanti.
- Un dato, nel secondo tentativo presentava, il valore 28, a fronte di un primo tentativo pari a 255. Presumendo un **errore di battitura**, è stato **modificato** in 280.

## Domande

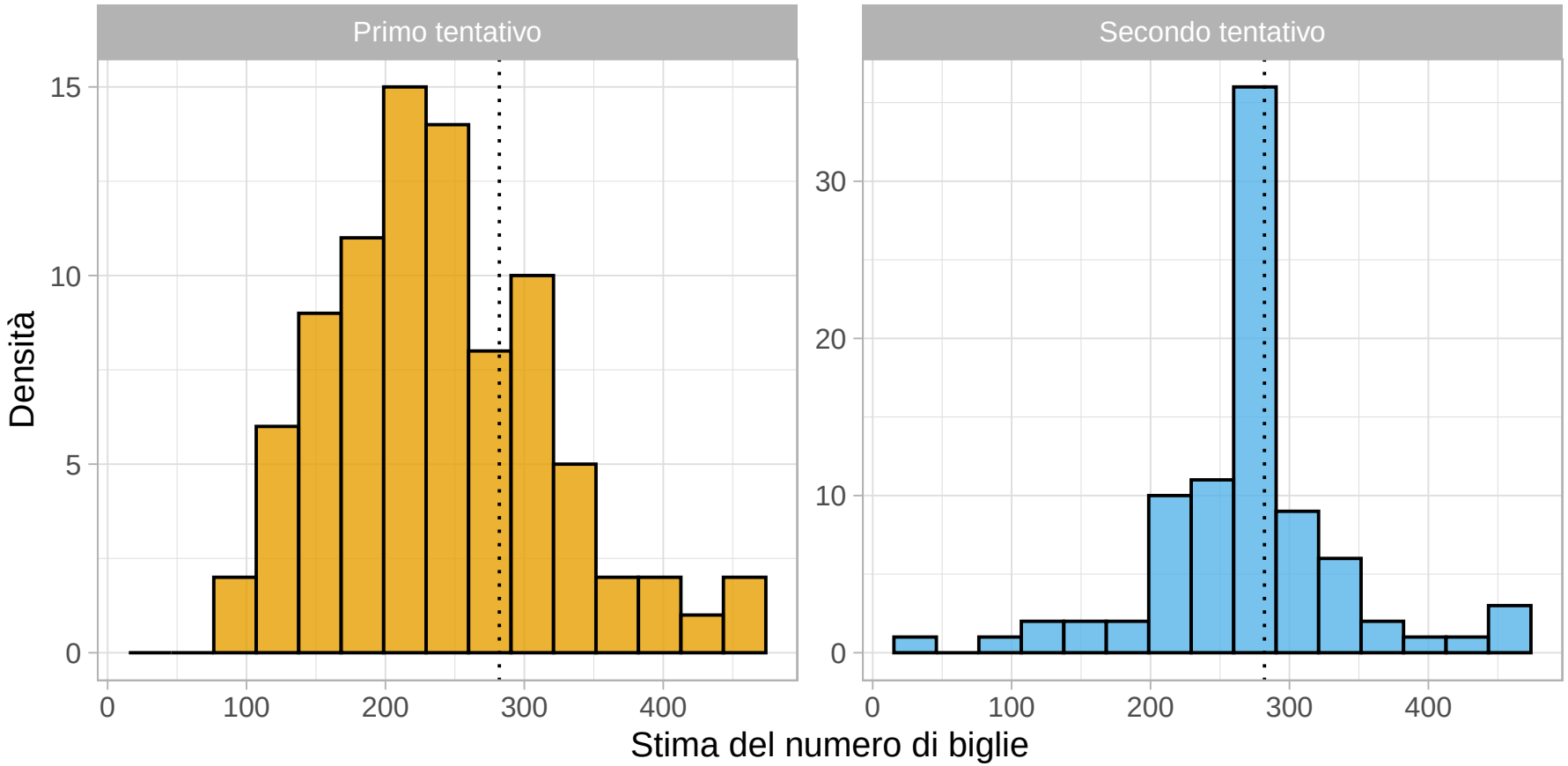
- Si descrivano le **principali caratteristiche** delle distribuzioni al primo ed al secondo tentativo.
- La **vox populi** si è rivelata affidabile?
- Ci sono **differenze** apprezzabili tra i due tentativi?

# Istogrammi




- È presente un **valore anomalo**, ovvero la coppia di dati $(670, 1096)$. Si è deciso, in maniera abbastanza arbitraria, di **escluderlo** da tutte le analisi successive.

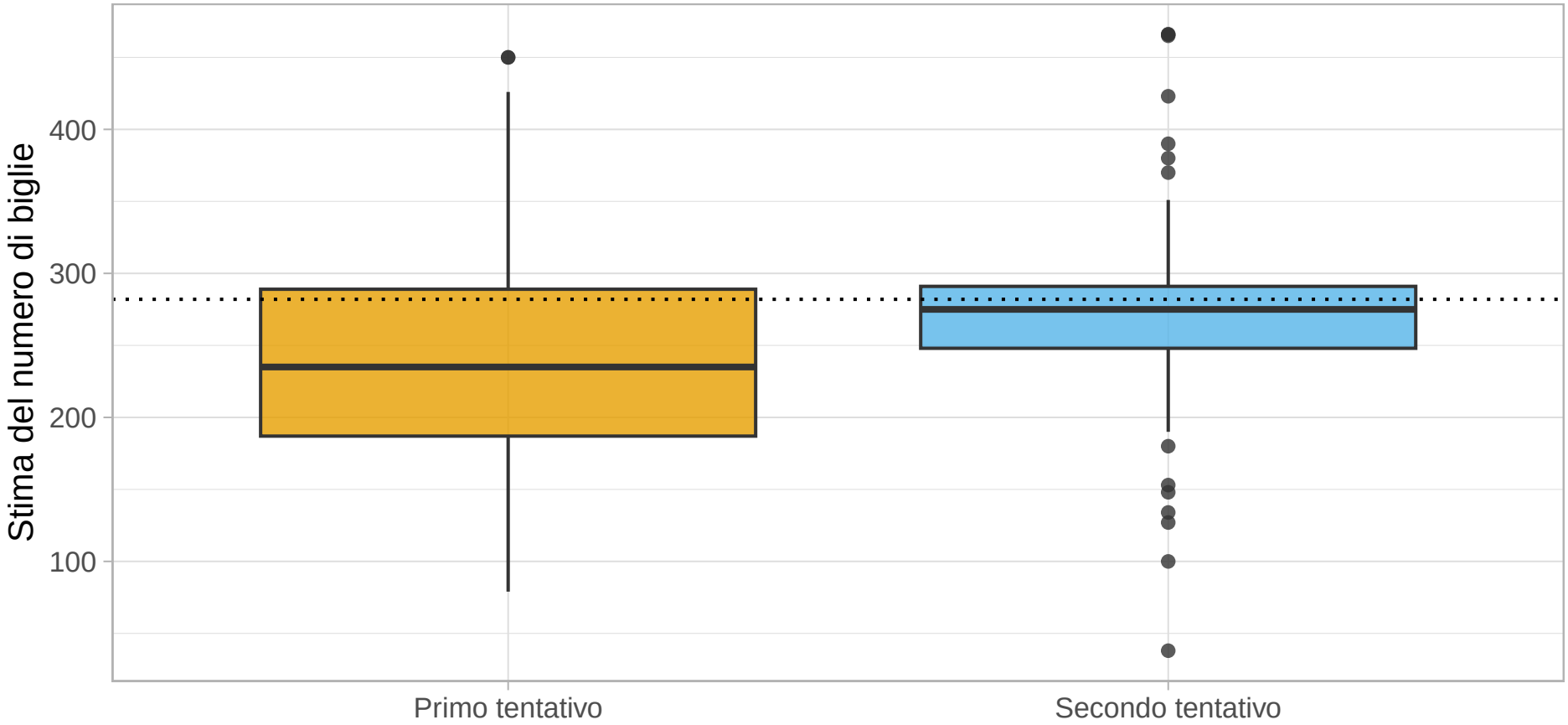
# Istogrammi




## La soluzione

- Il numero di biglie contenute nella bottiglia è **282**.

# Boxplot


Stima del numero di biglie
400
300
200
100
Primo tentativo
Secondo tentativo

# Statistiche descrittive

| | Primo tentativo | Secondo tentativo |
|---|---|---|
| Minimo | 79 | 38 |
| Primo quartile | 187 | 247 |
| Media | 239.05 | 269.74 |
| Mediana | 235 | 275 |
| Terzo quartile | 291 | 293 |
| Massimo | 450 | 466 |
| Varianza | 6101.72 | 4790.29 |
| Scarto quadratico medio | 78.11 | 69.21 |
| Scarto interquartile | 104 | 46 |
| MAD | 50 | 23 |
| Asimmetria di Pearson $\gamma$ | 0.54 | 0.10 |
| Asimmetria di Bowley | 0.08 | -0.22 |
| Curtosi $\kappa$ | 3.19 | 5.35 |

# Misurare gli errori

- Gli indici di variabilità misurano quanto i dati, le stime in questo caso, differiscono tra loro. In particolare, varianza e MAD misurano la distanza **dal proprio centro**.
- Purtroppo, il "centro" del primo tentativo è "**sbagliato**", perchè sottostima il numero corretto di biglie, cioè 282. Diremo che è una stima **distorta**.
- Di conseguenza, non possiamo usare la varianza per misurare la precisione.
- Siano $x_1, \ldots, x_n$ le stime di un generico tentativo. Due indicatori appropriati sono invece:

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - 282)^2, \qquad \frac{1}{n}\sum_{i=1}^{n}|x_i - 282|,$$

che prendono il nome di **errore quadratico medio** ed **errore assoluto medio**.

| | **Primo tentativo** | **Secondo tentativo** |
|---|---|---|
| Errore quadratico medio | 7946.77 | 4940.70 |
| Errore assoluto medio | 74.06 | 45.90 |

# Commento ai risultati

- Il numero di biglie presente nella bottiglia era 282. La folla è stata in grado di predirlo?
- La *vox populi* è stata **discretamente** affidabile nel primo tentativo **molto** affidabile nel secondo. Usando la mediana, la folla ha rispettivamente previsto **235** e **275** biglie.
- Inoltre, ci sono delle importanti **differenze** tra i due tentativi.
- Al primo tentativo, le stime sono in parte **distorte**, cioè si concentrano su un valore non corretto. Viceversa, dopo le informazioni aggiuntive, questa distorsione quasi scompare.
- Inoltre, la variabilità delle stime si **riduce** al secondo tentativo, come indicato ad esempio dall'**errore quadratico medio**.
- Il secondo tentativo manifesta una distribuzione **leptocurtica** ($\kappa > 3$), che indica la presenza di stime poco precise nonostante metà dei dati sia compreso tra $(247, 293)$.