Class 2c: Review of concepts in Probability and Statistics

Business Forecasting

Summarizing Data

Summary Statistics

Measures of Dispersion

- How much variation there is in the data?

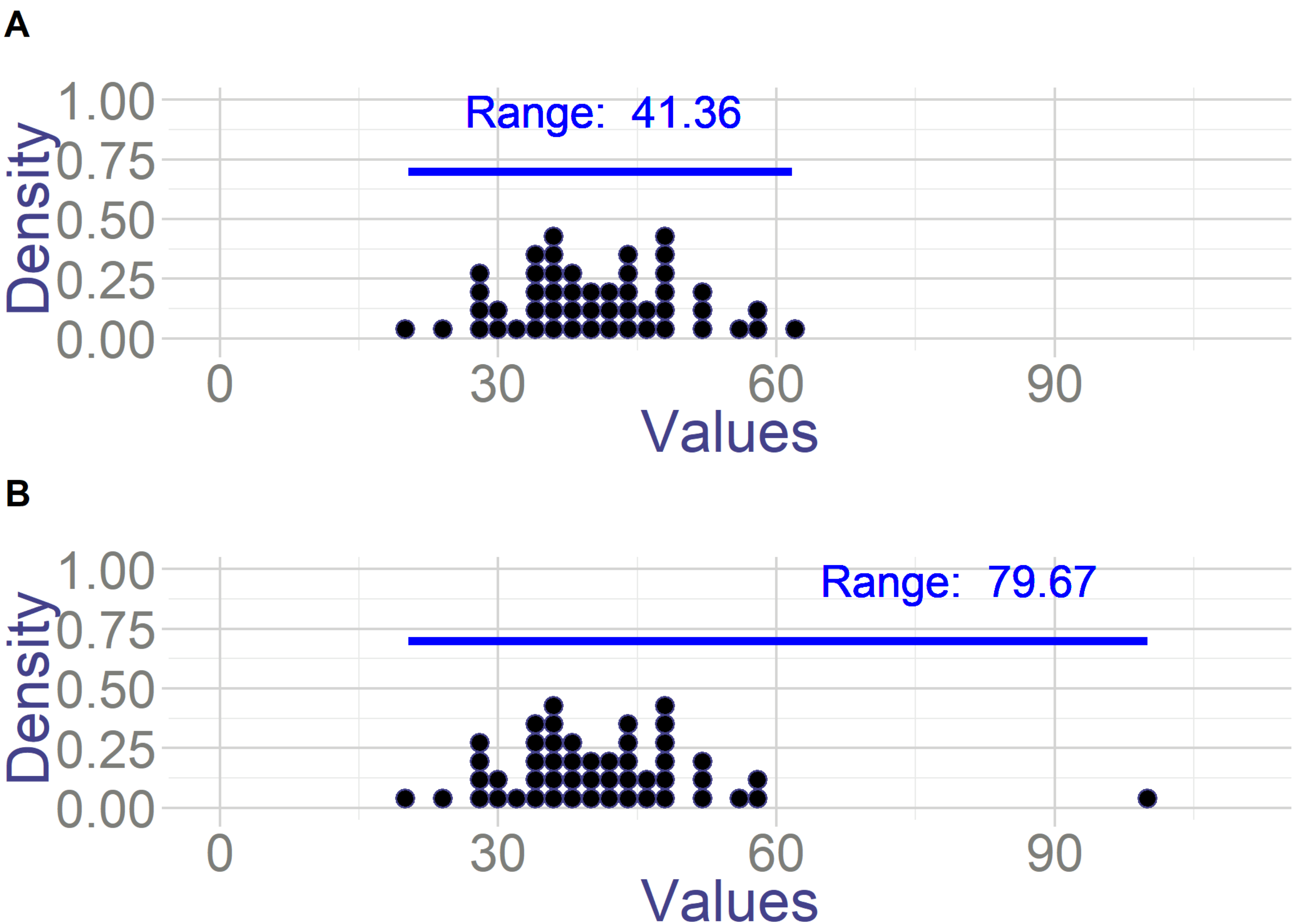
Range

- **Range** the difference between minimum and maximum value in the data

$$R = x_{max} - x_{min}$$

- What is the difference between the oldest and the youngest person with diabetes?
- **R**=77=97-20

- Very sensitive to outliers

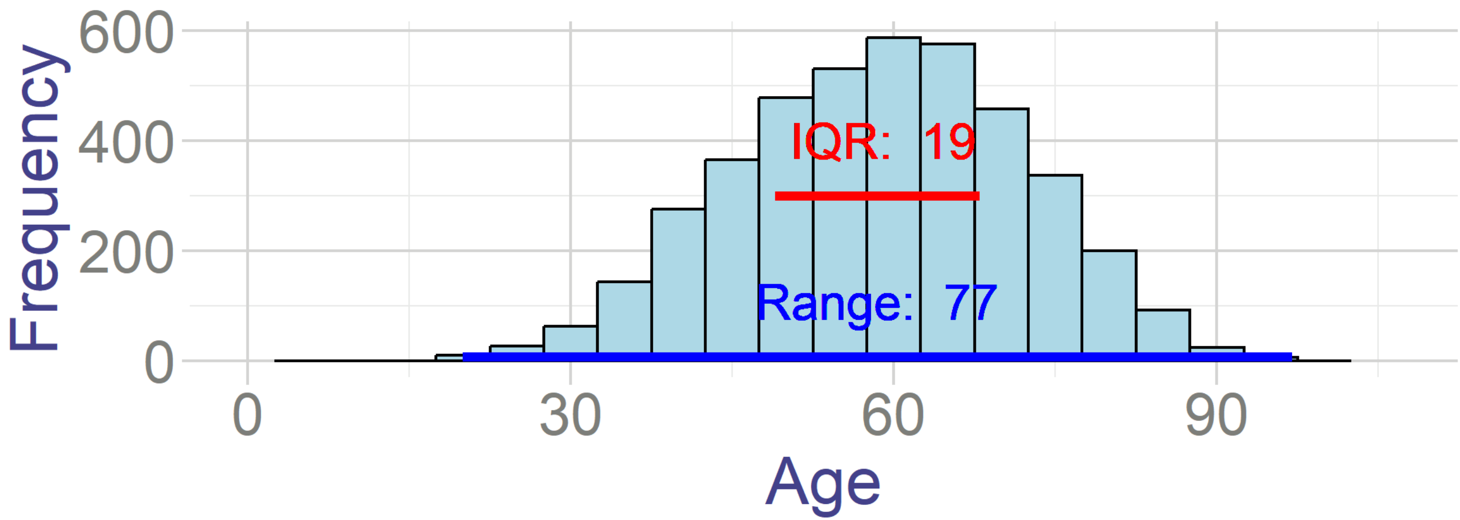


Interquartile Range

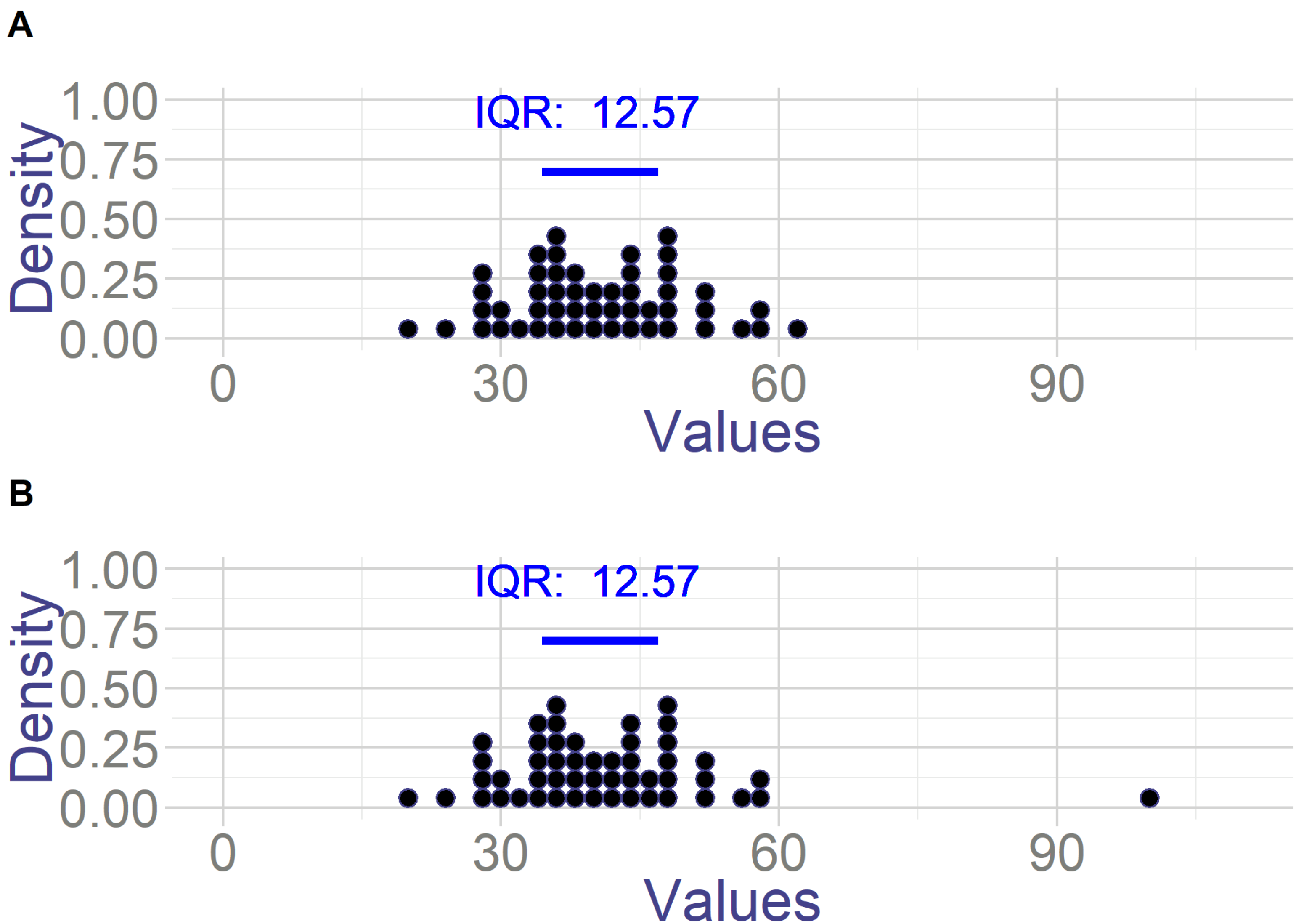
- **Interquartile range** is the difference between the first and the third quartile of the data:

$$IQR = q_3 - q_1$$

- What is the IQR of age in people with diabetes?
- **IQR**=19=68-49
- 50% of the sample is between q_3 and q_1



- Is it more or less sensitive to outliers than range?



Interquartile Range

Example with data

- What is the IQR?

Show 4 entries

VideoTitle	Views
TikTok Video 1	30
TikTok Video 2	17
TikTok Video 3	22
TikTok Video 4	24

Showing 1 to 4 of 20 entries

Previous 1 2 3 4 5 Next

Example with data

Here is a (smaller) data on distribution of how many views have various tik-tok videos.

- Suppose that all views triples and 1000 additional people viewed them as well

$$y_i = 3x_i + 1000$$

- What is new IQR?

Show 4 entries

VideoTitle	OldViews	NewViews
TikTok Video 1	30	1090
TikTok Video 2	17	1051
TikTok Video 3	22	1066
TikTok Video 4	24	1072

Showing 1 to 4 of 20 entries

Previous 1 2 3 4 5 Next

IQR

- Order of observations was not affected, so same observations correspond to the first and the third quartile

$$q_1^{New} = 3q_1^{Old} + 1000$$

$$q_3^{New} = 3q_3^{Old} + 1000$$

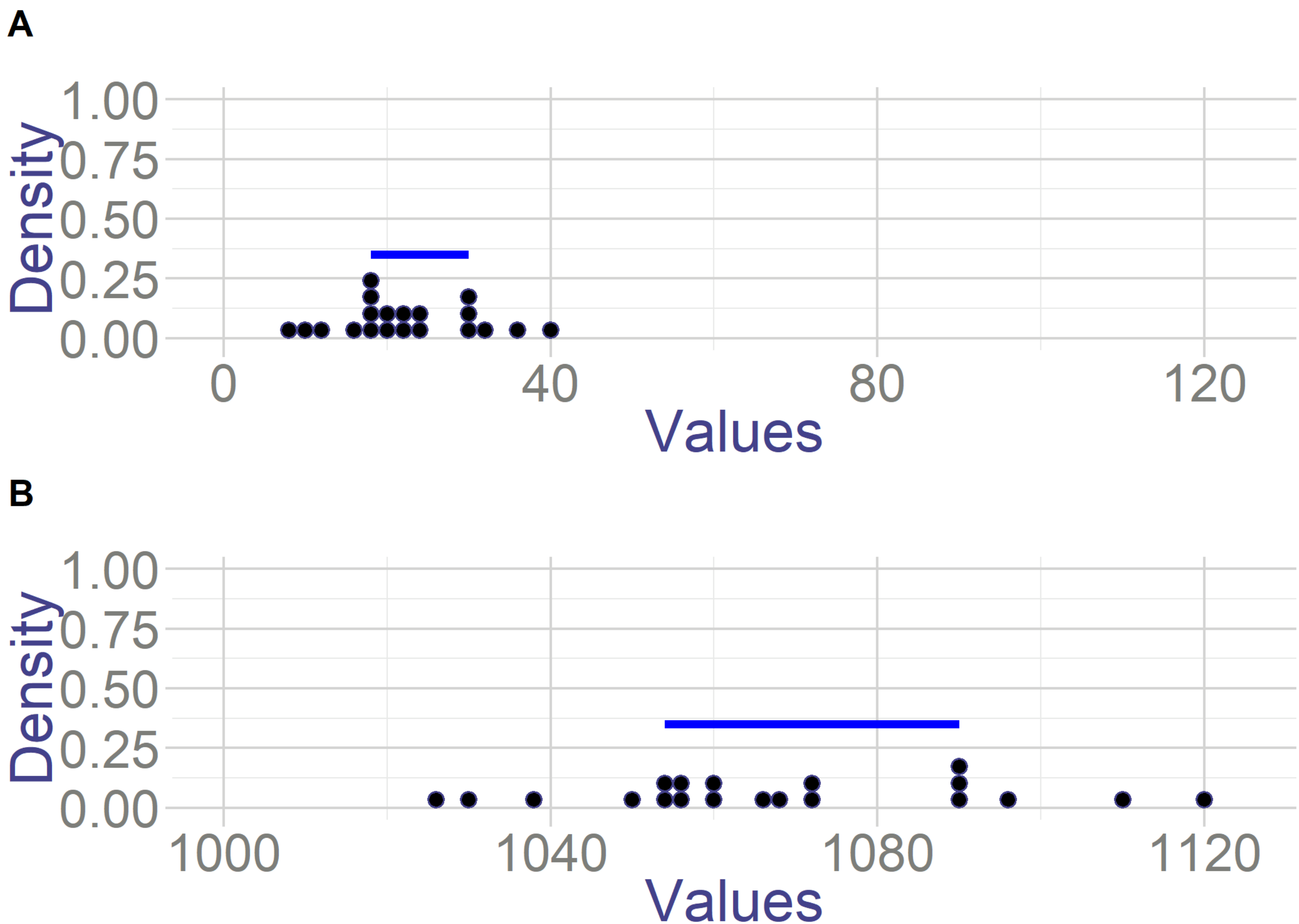
- And more generally, for

$$y_i = bx_i + a$$

$$v_p^y = bv_p^x + a$$

- So what does it mean for IQR?

$$IQR^{New} = q_3^{New} - q_1^{New} = 3q_3^{Old} - 3q_1^{Old} = 3 * IQR^{Old}$$


A
1.00
0.75
0.50
0.25
0.00
Density
0
40
80
120
Values
B
1.00
0.75
0.50
0.25
0.00
Density
1000
1040
1080
1120
Values

Variance & Standard Deviation

Variance measures how much observations deviate from the mean:

- **Population variance**:

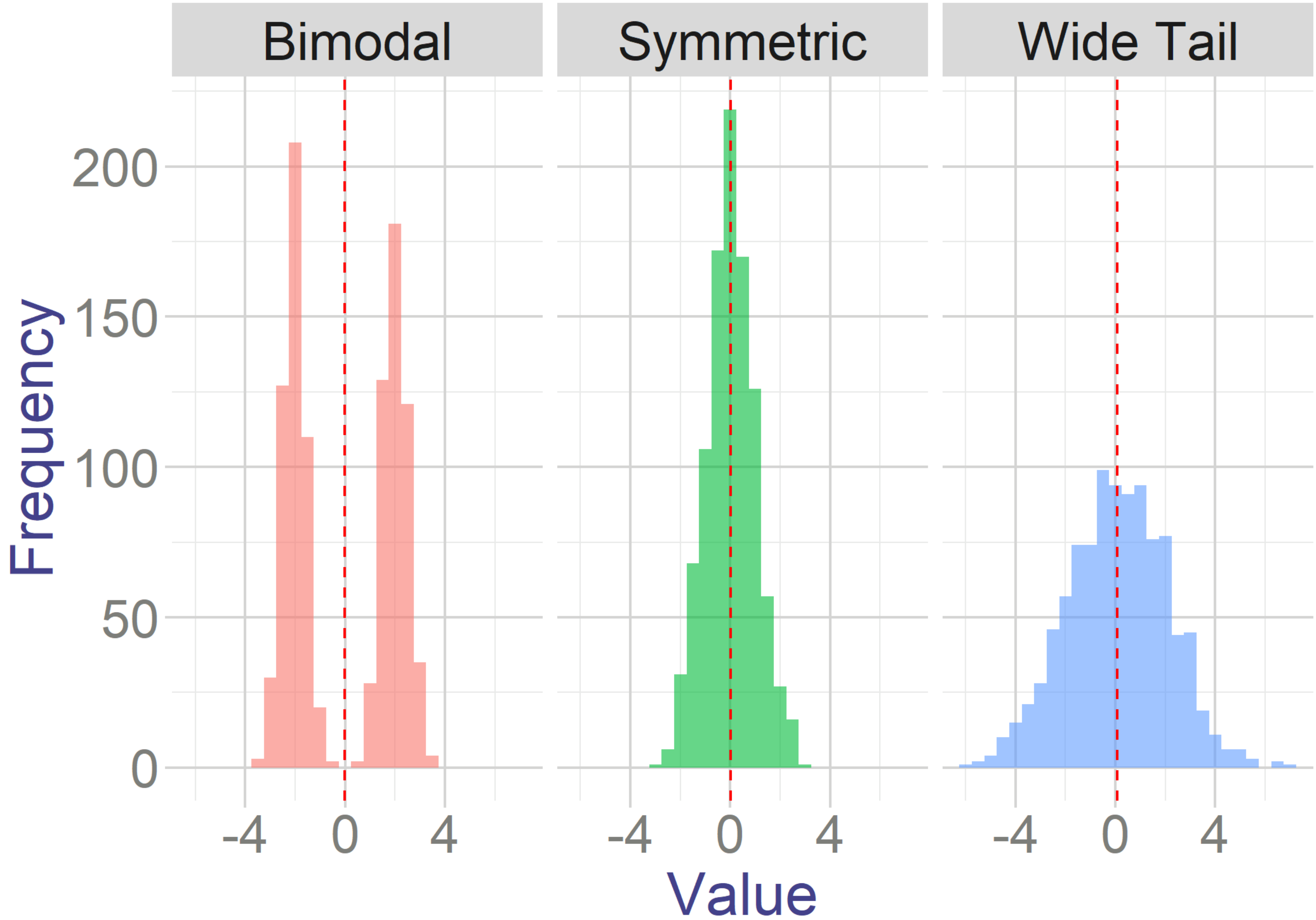
$$\sigma^2 = E[(X-\mu)^2] = \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2$$

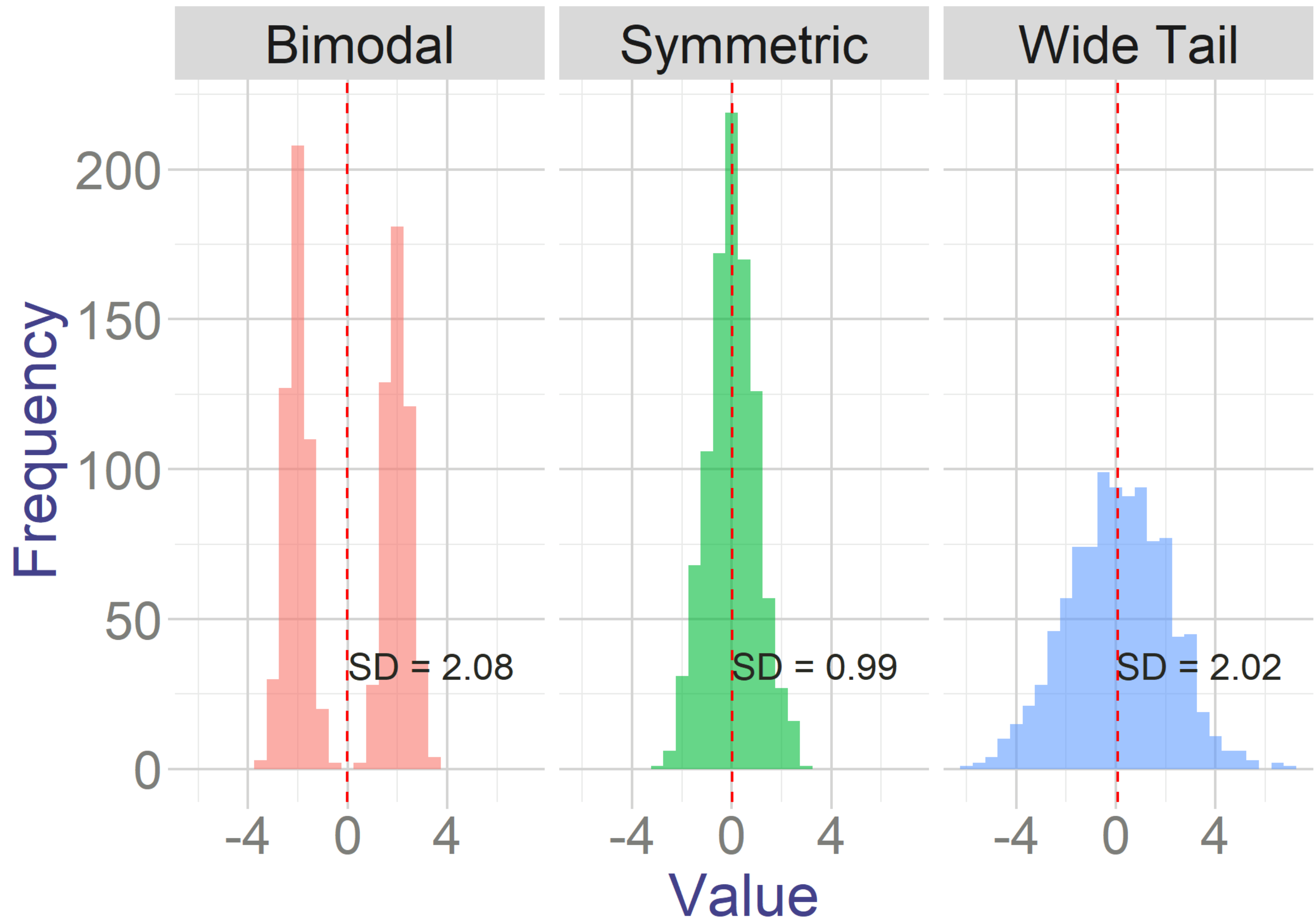
- But this does not have the right units...
- **Population standard deviation** deviation:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2}$$

- Why do we first take squares and then take square root?

- Can't we just do $\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)$?
- NO! Because $\sum_{i=1}^{N}(x_i-\mu)=0$


Bimodal
Symmetric
Wide Tail
Frequency
200
150
100
50
0
-4
0
4
Value


Bimodal
Symmetric
Wide Tail
Frequency
200
150
100
50
0
SD = 2.08
SD = 0.99
SD = 2.02
-4
0
4
Value

Sample equivalents

- **Sample variance**:

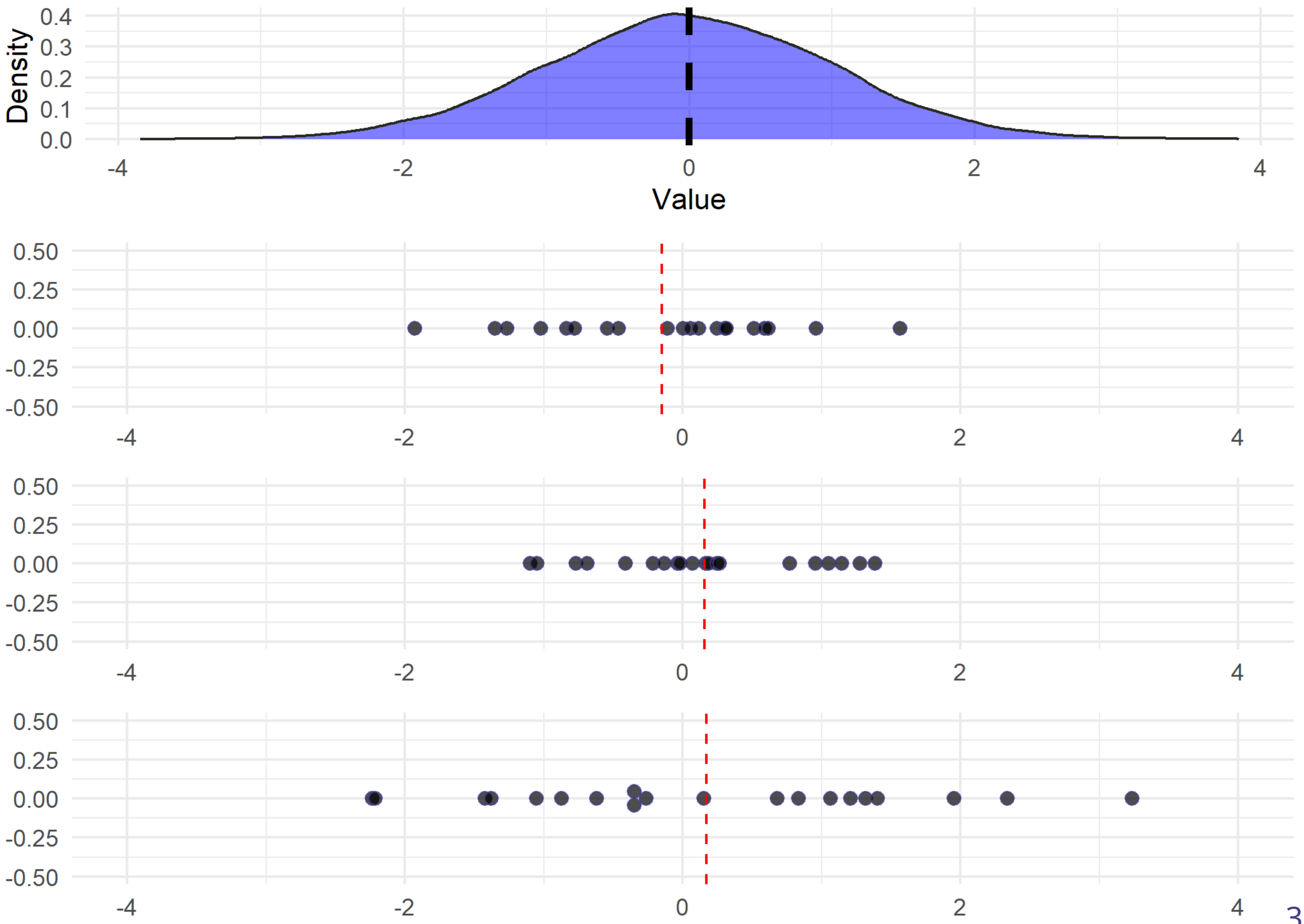
$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- **Sample standard deviation** deviation:

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- Why we divide by $n-1$ rather than n?

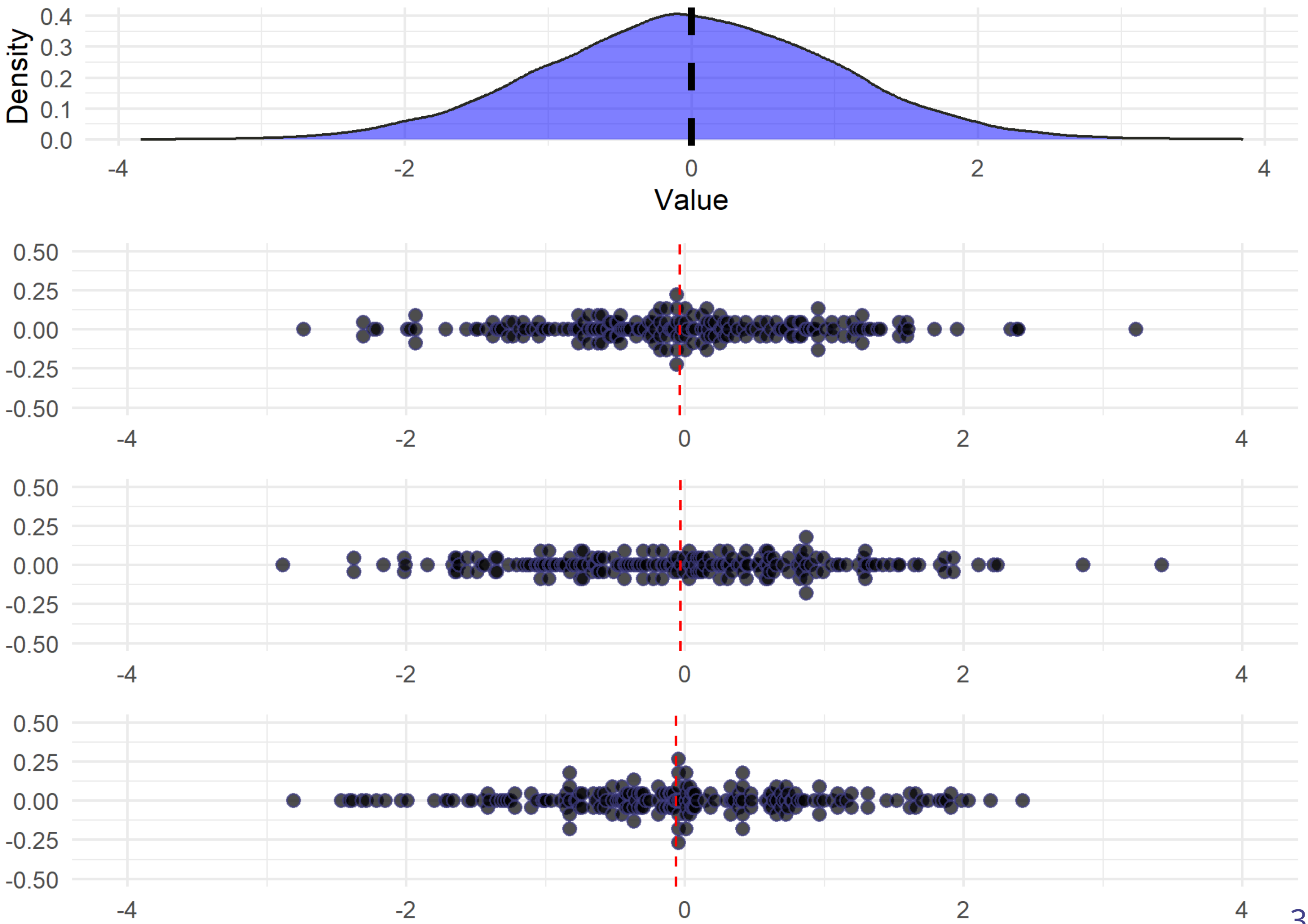
- **Intuition** - observed values usually fall closer to the sample mean than to the population mean



Sample equivalents

- So the deviations from the sample mean underestimate the population standard deviation
- So we divide by a smaller number to correct for it
- In big sample $\frac{1}{n}$ and $\frac{1}{n-1}$ are similar, so correction doesn't matter as much

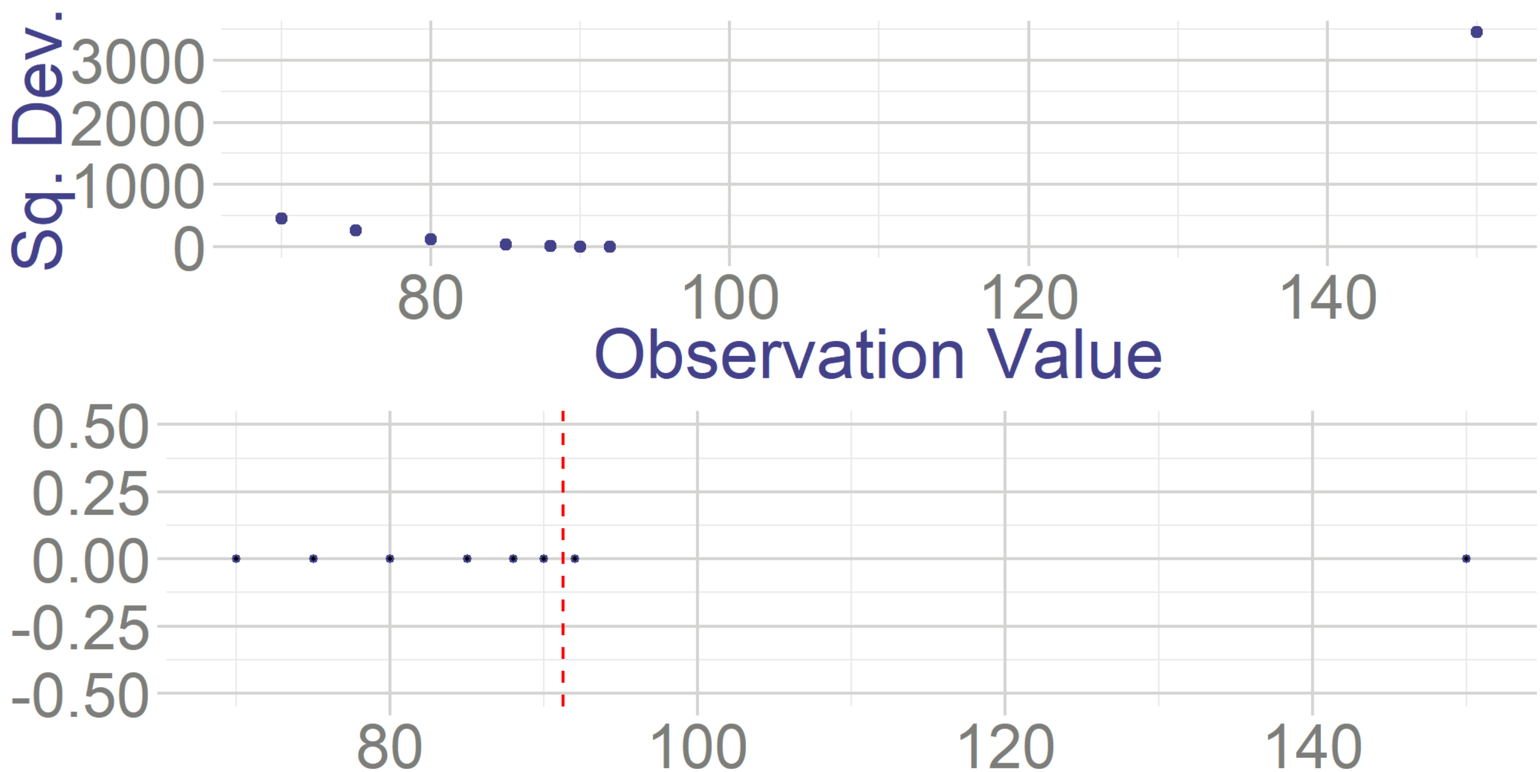
- **Intuition** - in big samples, our estimate of the population mean is already good, no need to correct



Sample equivalents

Are they robust to outliers?

- Very sensitive because squaring deviation amplifies large deviation more than small deviations.



Variance in the aggregated data

Can we calculate variance in the population from aggregate data?

Show 6 entries

Neighborhood	Average_Income	Population
Polanco	60000	10000
Condesa	45000	20000
Roma	35000	30000
Tepito	15000	5000
Coyoacán	30000	25000
Santa Fe	25000	18000

Showing 1 to 6 of 12 entries   Previous 1 2 Next

$$\frac{\sum_z N_z(\bar{x}_z - \mu)^2}{\sum_z N_z} \neq \sigma^2$$

- **Intuition:** we are missing the variation within the neighborhoods

Variance in the aggregated data

- We can't do it if there is variation within groups
- But we can if there is no variation within groups
  - *Example*: frequency distribution of discrete variables

Show 6 entries

Age	n_i	p_i
20	4	0.001
21	2	0
22	4	0.001
23	3	0.001
24	5	0.001
25	7	0.002

Showing 1 to 6 of 78 entries

Previous 1 2 3 4 5 ... 13 Next

- For simplicity, assume we have data on all population:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2$$

$$= \frac{(20-\mu)^2+(20-\mu)^2+(20-\mu)^2+\ldots}{N}$$

$$= \frac{4\cdot(20-\mu)^2+2\cdot(21-\mu)^2+\ldots}{N}$$

$$= \frac{\sum_a N_a(\bar{x}_a-\mu)^2}{\sum_a N_a}$$

- Where N_a is number of people with age a

Coefficient of Variation

Coefficient of Variation divides the standard deviation by the mean.

$$C.V. = \frac{\sigma}{|\mu|}$$

And sample eqiuvalent

$$c.v. = \frac{s}{|\bar{x}|}$$

- Why?
  - It expresses standard deviation as proportion of the mean
    - Small value means variation is low compared to the mean
  - It is unit free
  - You can compare it across variables with different units/magnitudes

Coefficient of Variation

Example - variation of stocks in different currencies

Show 6 entries

Date	MXN_Stock	USD_Stock
2023-07-01	91.59	1.01
2023-07-02	96.55	1.16
2023-07-03	123.38	1.02
2023-07-04	101.06	1.07
2023-07-05	101.94	1.09
2023-07-06	125.73	0.9

Showing 1 to 6 of 20 entries

Previous 1 2 3 4 Next

- **Standard deviation**:
  - USD: 0.149
  - MXN: 14.59
- **Coefficient of variation**:
  - USD: 0.12
  - MXN: 0.14

Coefficient of Variation

So more generally, if $y_i = bx_i$, then

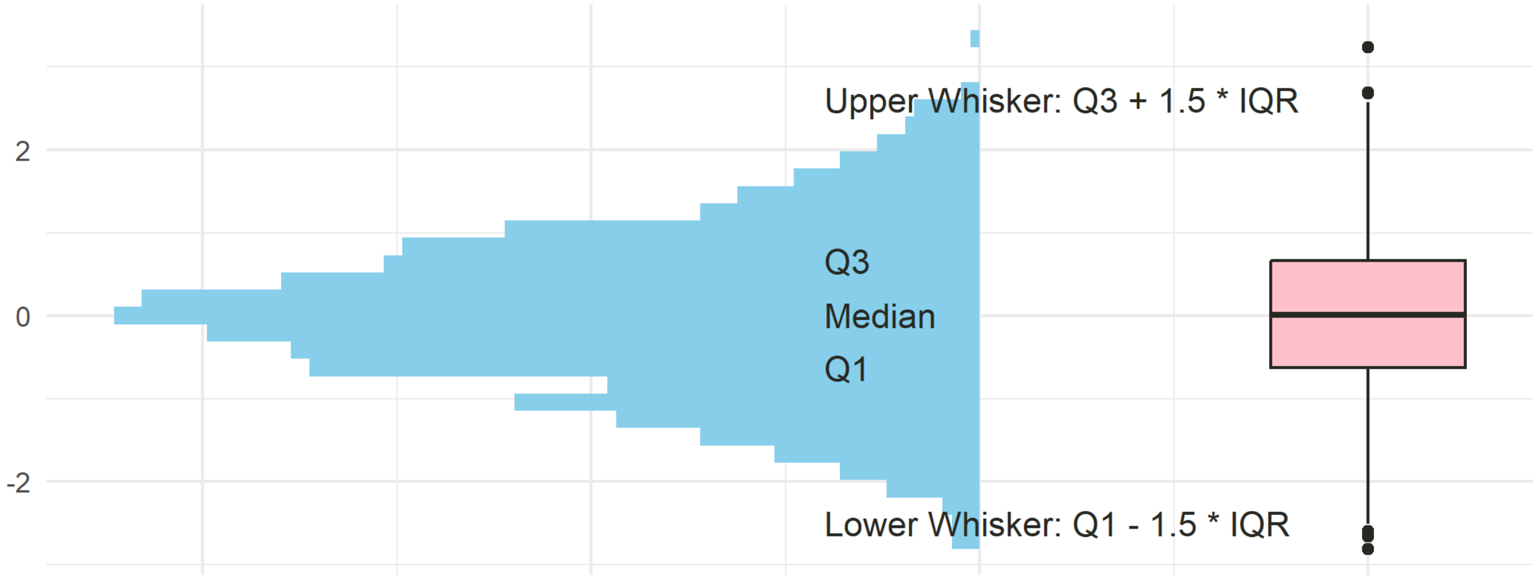
$$C.V._y = \frac{\sigma_y}{|\mu_y|} = \frac{|b|\sigma_x}{|b\mu_x|} = C.V._x$$

What if $y_i = bx_i + a$? Then

$$C.V._y = \frac{\sigma_y}{|\mu_y|} = \frac{|b|\sigma_x}{|b\mu_x + a|} \neq C.V._x$$

Box and Whiskers plot

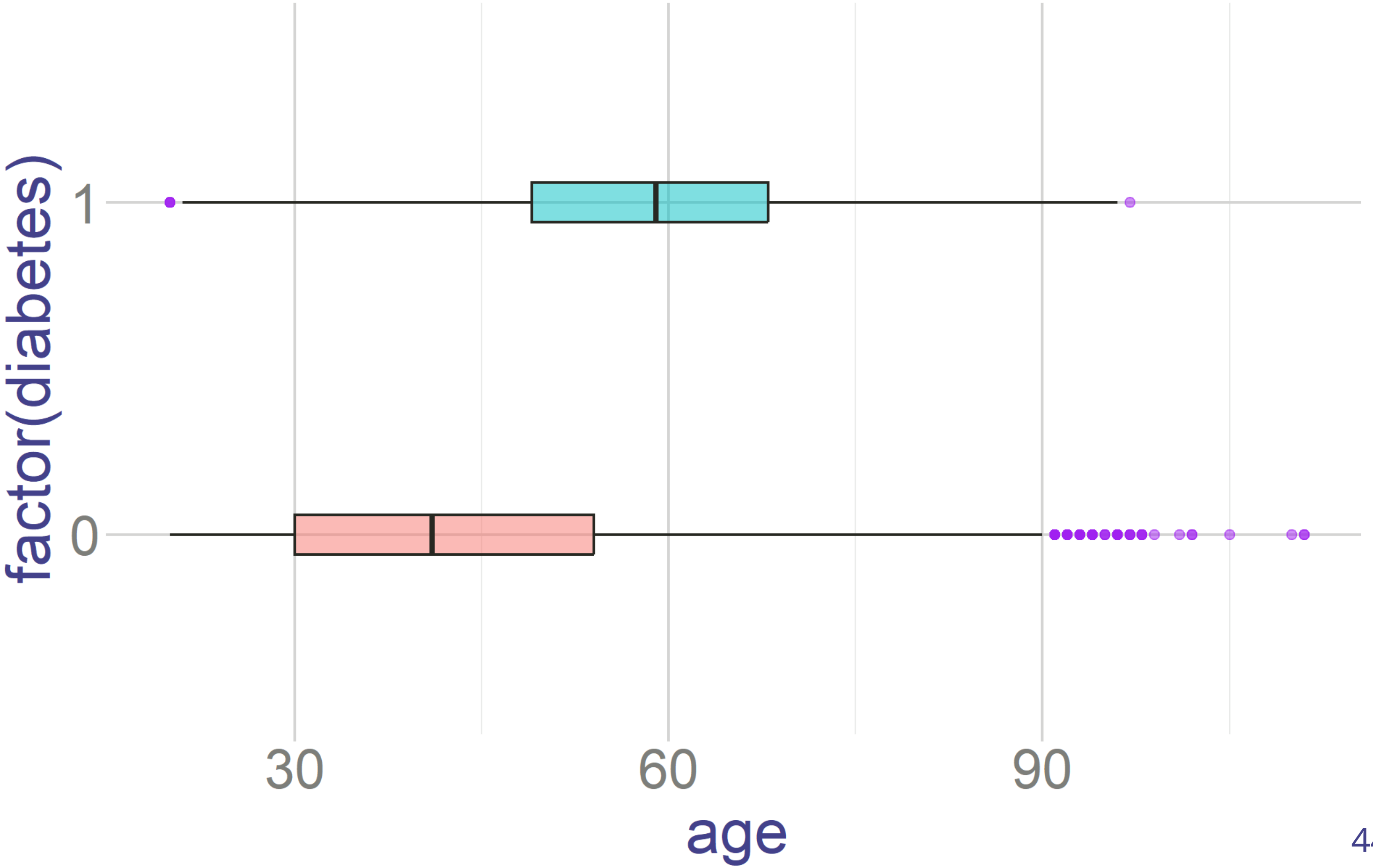
- Helps to see the distribution of the data
- Helps to see to see the outliers
  - Outliers are useful to see anomalies and potential errors in data colection



Box and Whiskers plot

Dataset comparisons

- They summarize data very well



Box and Whiskers plot

Dataset comparisons

- They summarize data very well