Class 2c: Review of concepts in Probability and Statistics

Business Forecasting

Summarizing Data

Summary Statistics

Measures of Dispersion

- Suppose a store has an average daily revenue of 10 000 pesos
- It could be that on each day it has exactly 10 000 pesos revenue
- Or it could be that on half of days it gets 20 000 pesos but on other half it gets 0 pesos
- Dispersion makes a big difference, especially when trying to understand risk!

Range

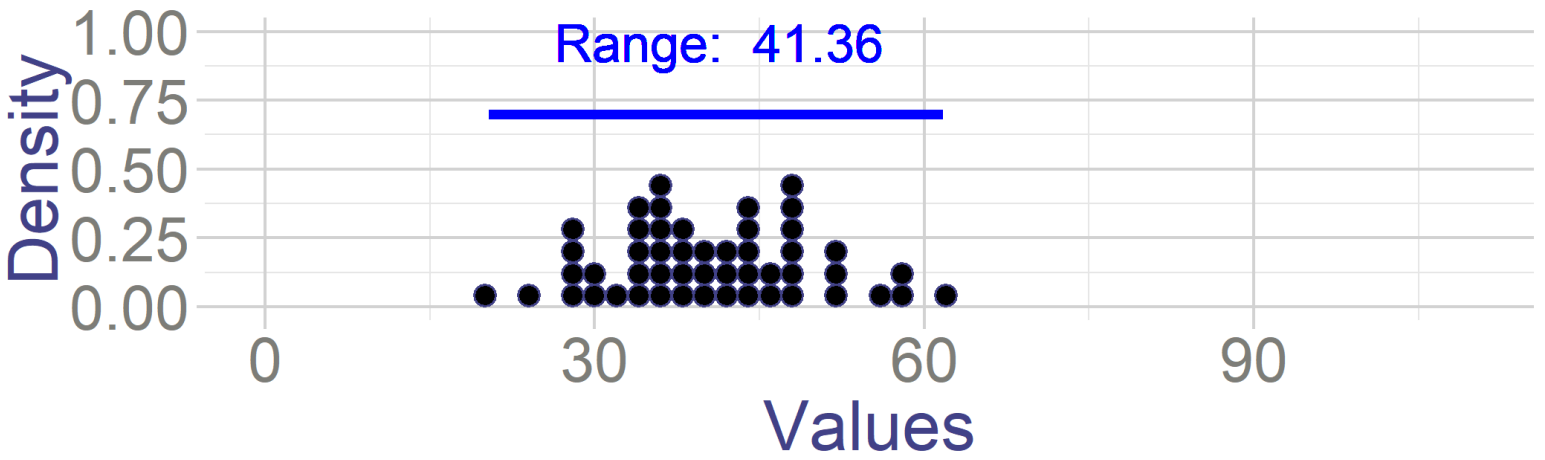
- **Range** the difference between minimum and maximum value in the data

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

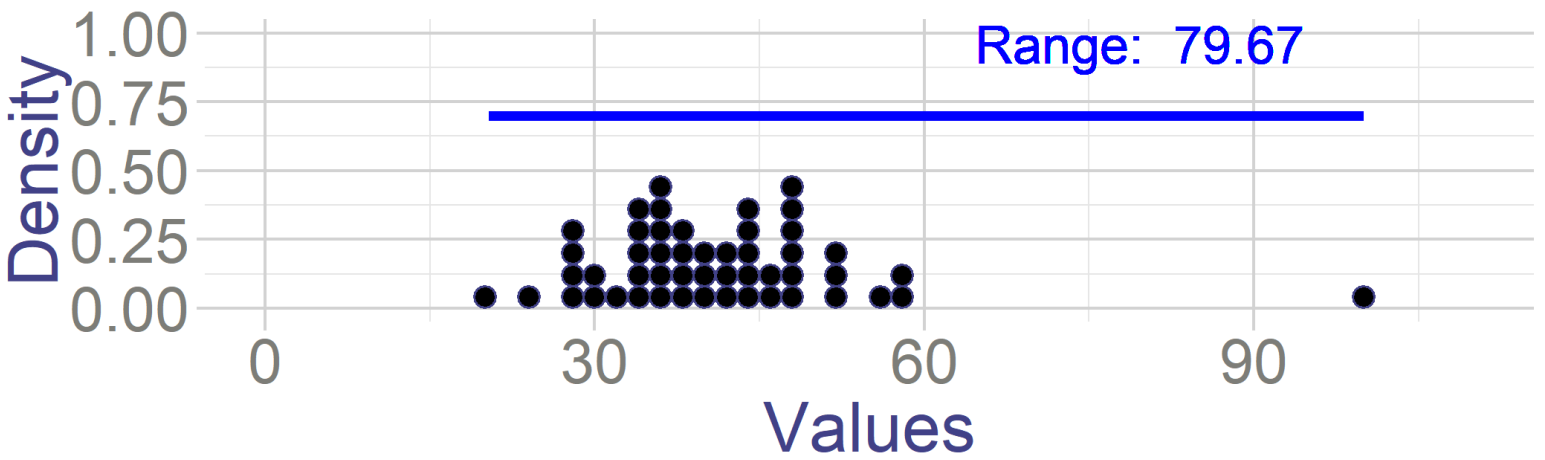
- What is the difference between the oldest and the youngest person with diabetes?
- **R**=77=97-20

- Very sensitive to outliers

A



B

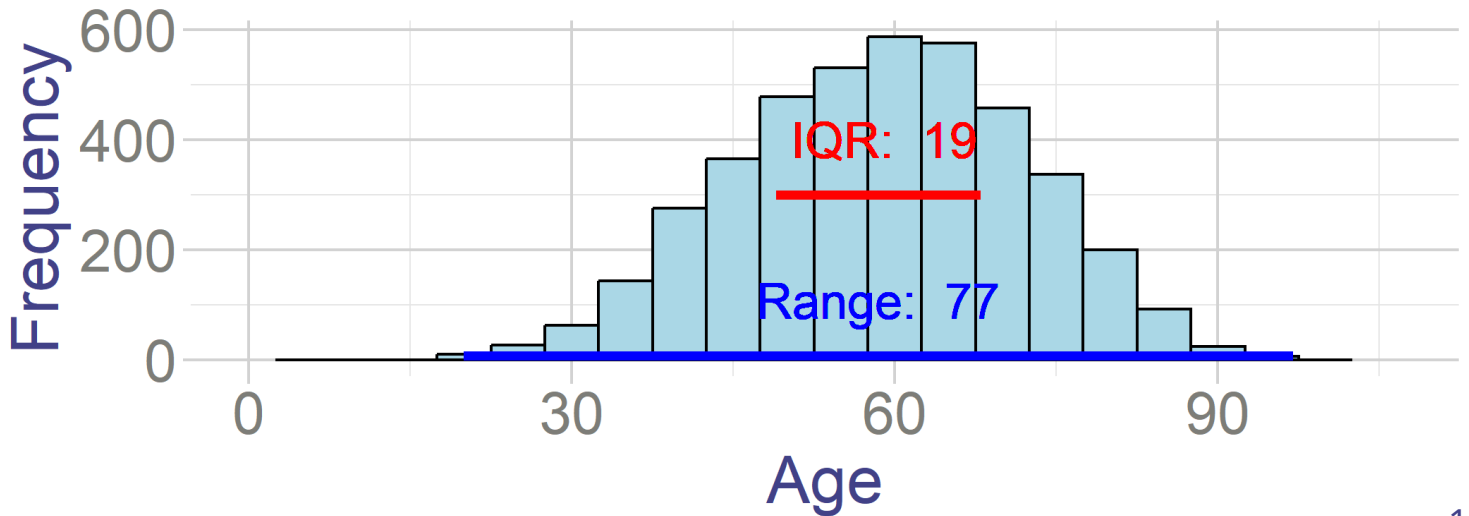


Interquartile Range

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

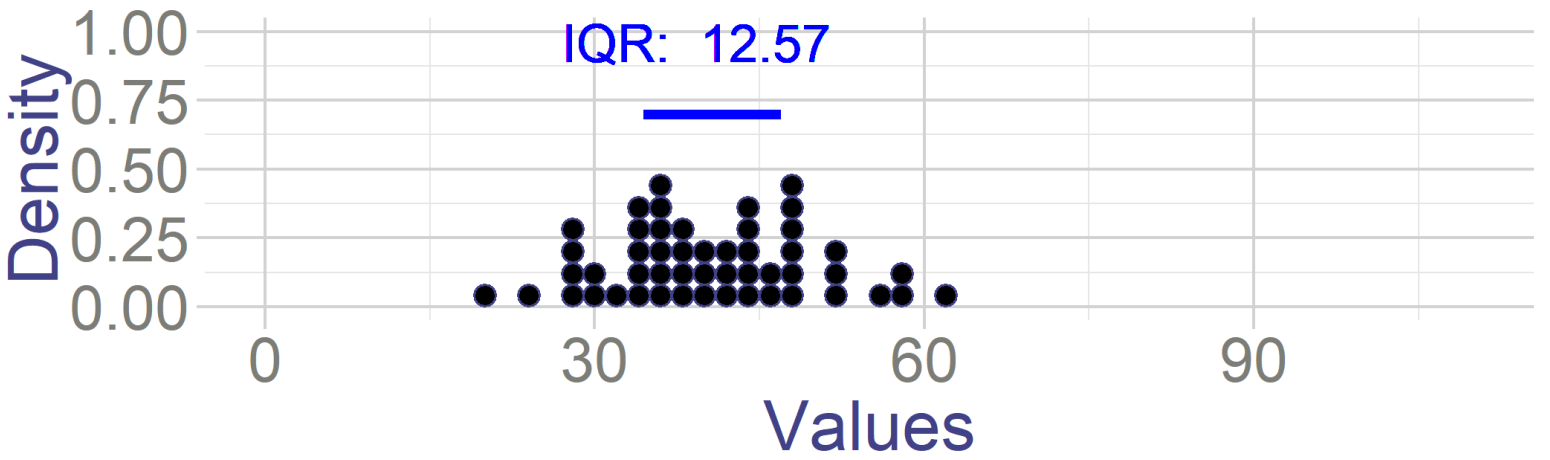
$$IQR = q_3 - q_1$$

- 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

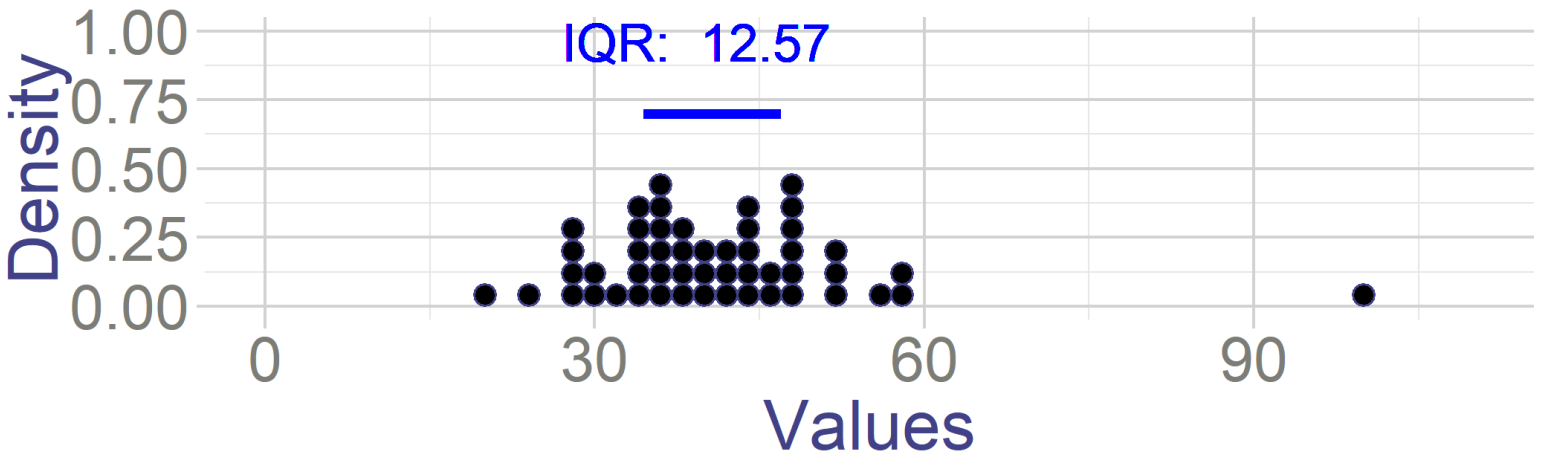


- Is it more or less sensitive to outliers than range?

A



B



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

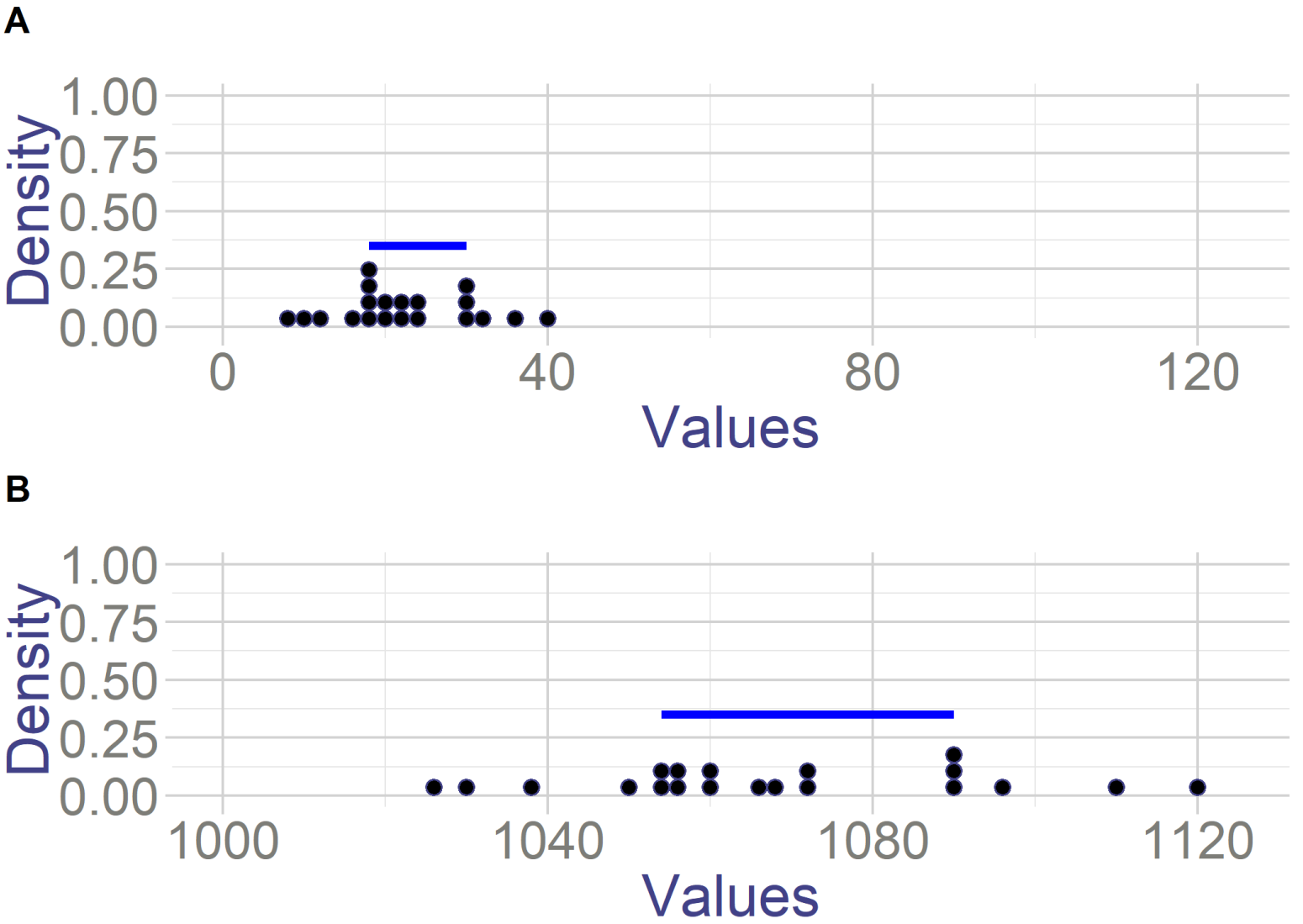
$$y_i = bx_i + a$$

and $b > 0$

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

- if $b < 0$ then the order reverses.
- 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
Density
1.00
0.75
0.50
0.25
0.00
0
40
80
120
Values
B
Density
1.00
0.75
0.50
0.25
0.00
1000
1040
1080
1120
Values

Variance & Standard Deviation

Variance measures how far an average observation is from the mean:

- **Population variance**:

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

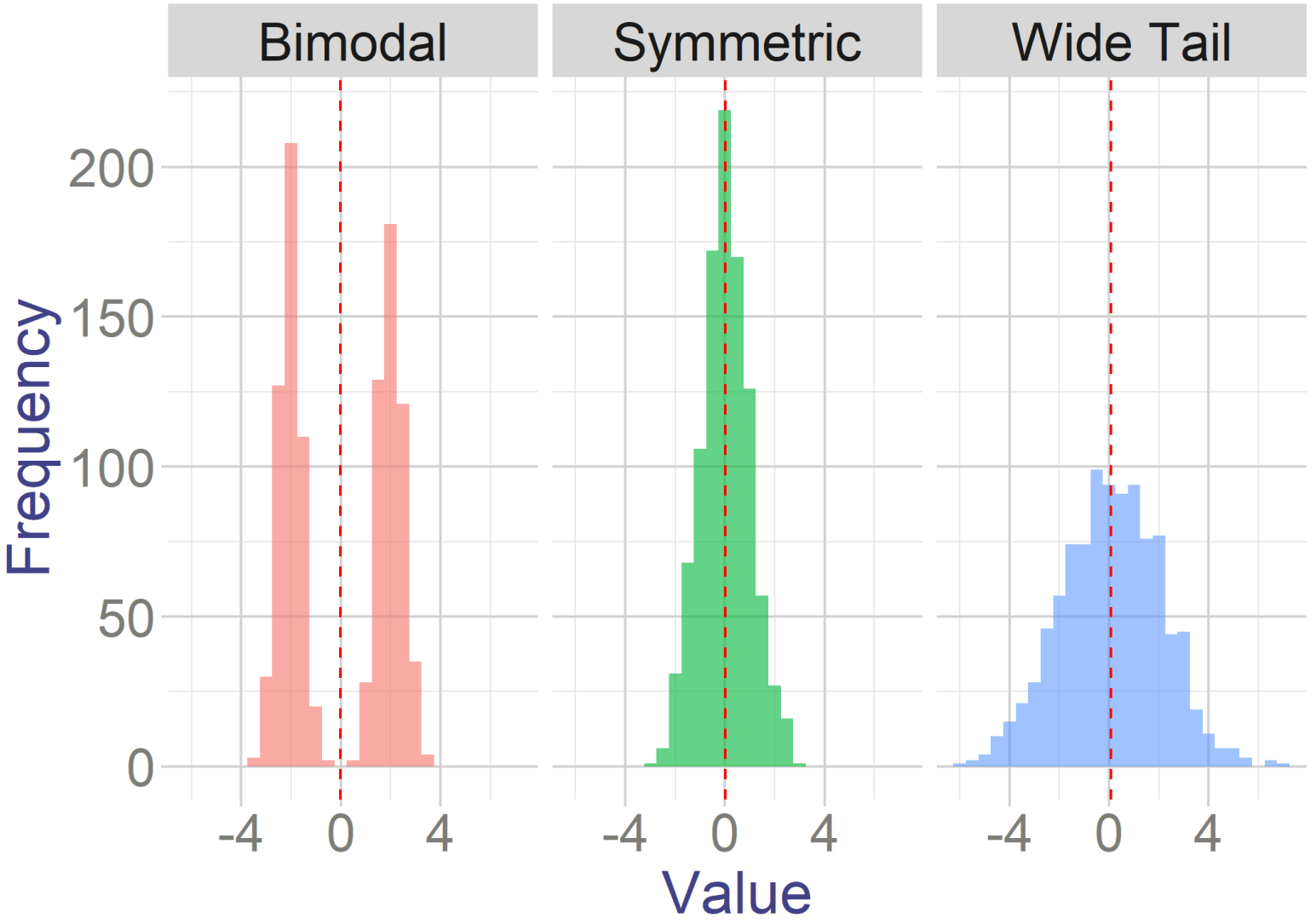
  For Discrete Variables it can be: $\sigma^2 = \sum_k P(X=k)(k-\mu)^2$, where k is any possible value that X can take.
- But variance does not have the right units since it squares everything...
- **Population standard deviation** deviation:

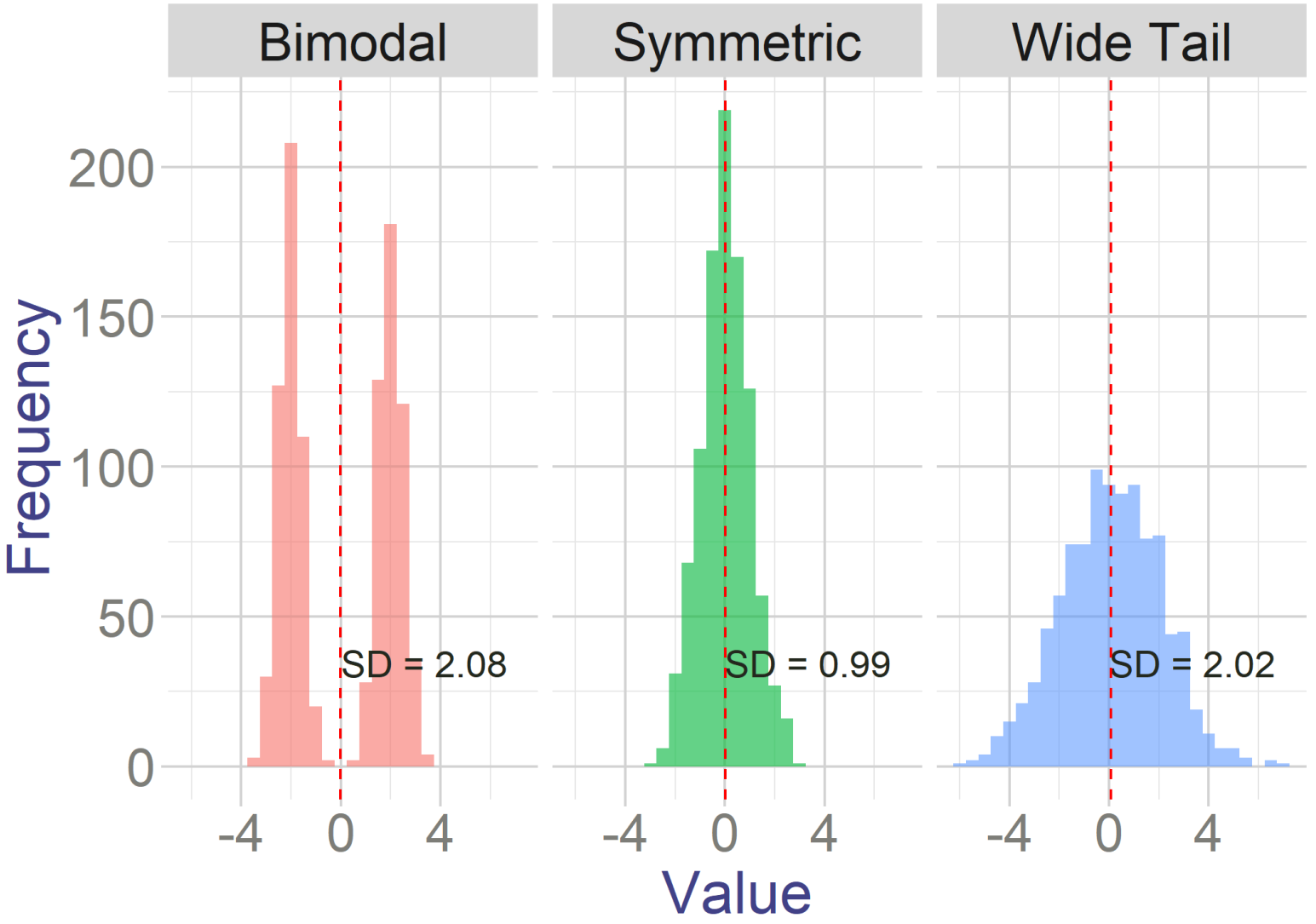
$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2}$$

Variance & Standard Deviation

- 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
  $sum_{i=1}^{N}(x_i - \mu) = 0$
- Why don't we just do Mean Absolute Deviation?
  - $MAD = frac1N \sum_{i=1}^{N} |(x_i - \mu)|$
  - MAD is not differentiable at 0 :(
- You can use it, but it's a different measure and will give different numbers. Why?
- Variance puts more weight on far away observations.
- It's a weighted average distance, where weights are distances themselves.

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


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

Variance & Standard Deviation

Consider two bets/situations:

- Bet A: with 75% you get 200 pesos and with 25% you lose me 200 pesos
    - 75% chance your life goes normal and you keep making money
    - 25% change your house burns down
- Bet B: with 75% you get 110 pesos and with 25% you get 70
    - When your life goes normal you get 110 (you pay 90 for insurance)
    - When your house burns down you are paid some compensation (70)
- Compute expected value and variance of each bet
- Which one would you prefer?
- What if I change Bet B:
    - Bet B: with 75% you get 109 pesos and with 25% you get 69
- That's how insurance companies make profits

Sample equivalents

- **Sample variance**:

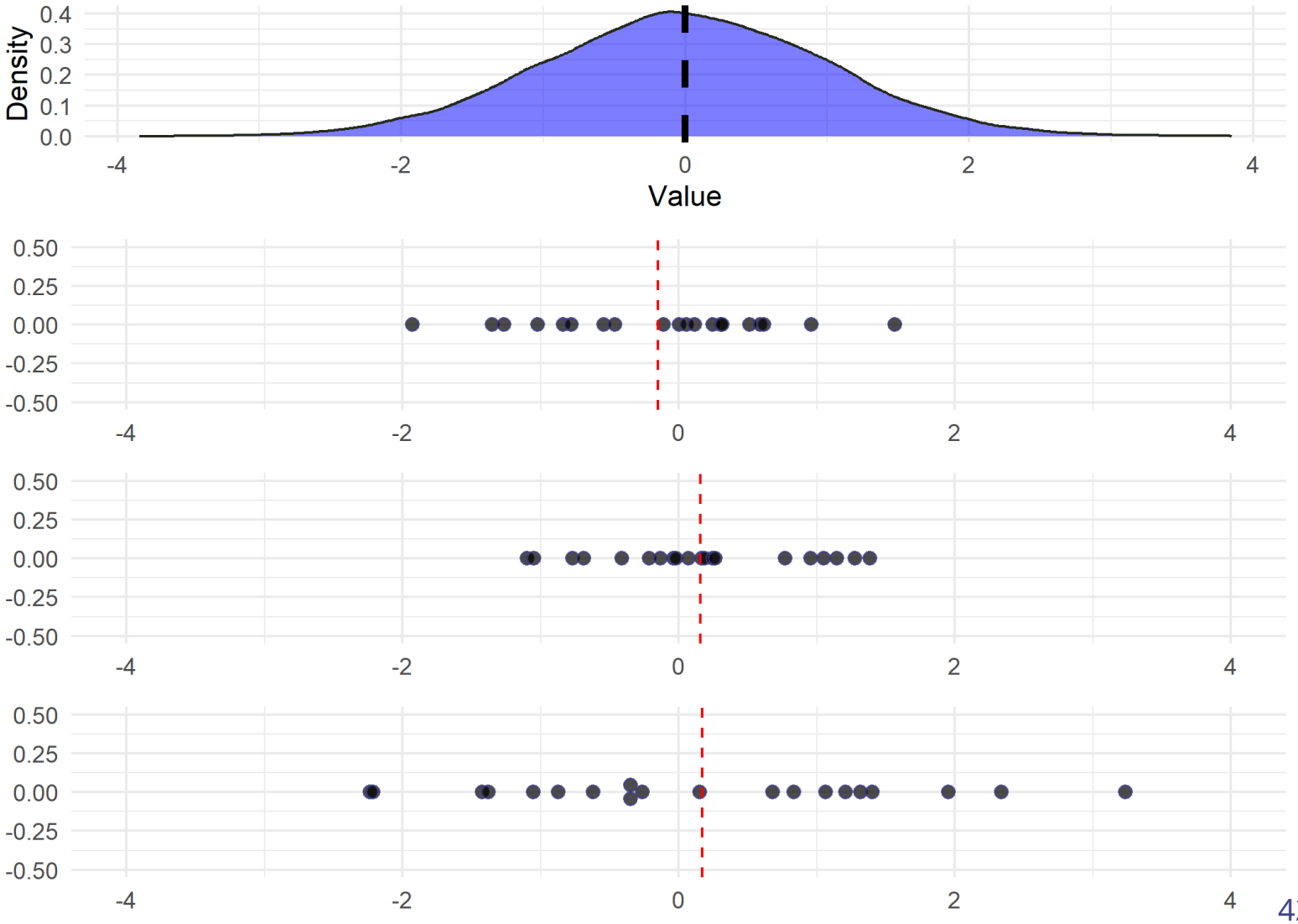
$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2)$$

  - **Sample standard deviation** deviation:

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

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

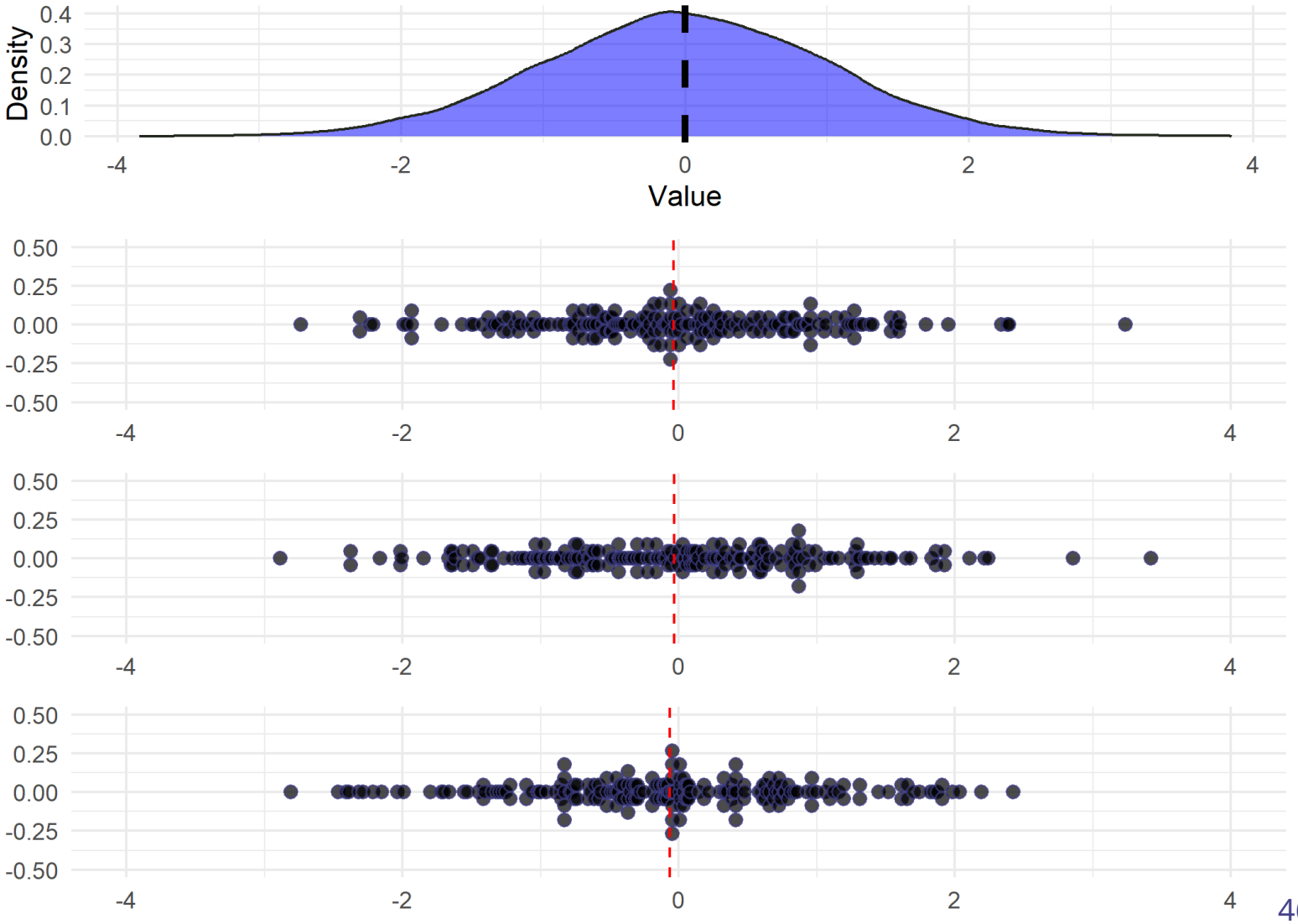
- **Intuition** - observed values usually fall closer to the sample mean than to the population mean. Distances are artificially small.



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

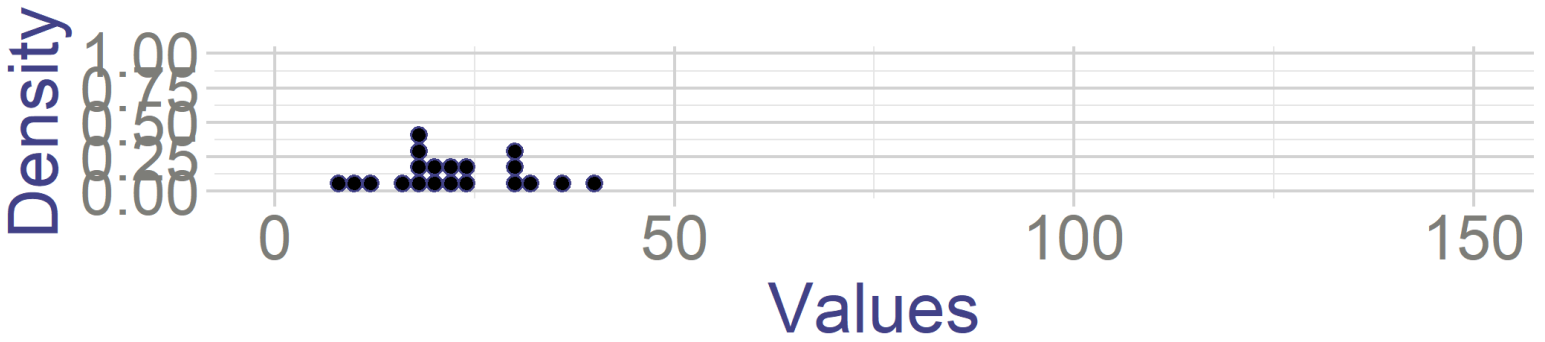
- **Intuition** - in big samples, our estimate of the population mean is already good, no need to correct



Standard Deviation

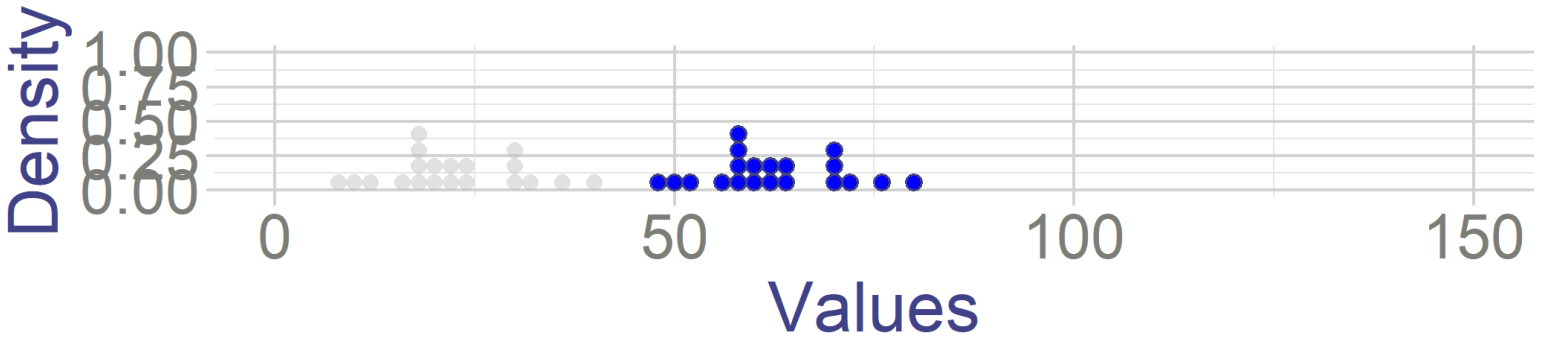
Consider a random variable X with $E(X) = \mu_x$ and standard deviation σ_x.

Ex: X is number of instagram followers distributed like this:



Standard Deviation

- What happens to mean and standard deviation if everyone gets 40 more followers?
- $Y = X + 40$



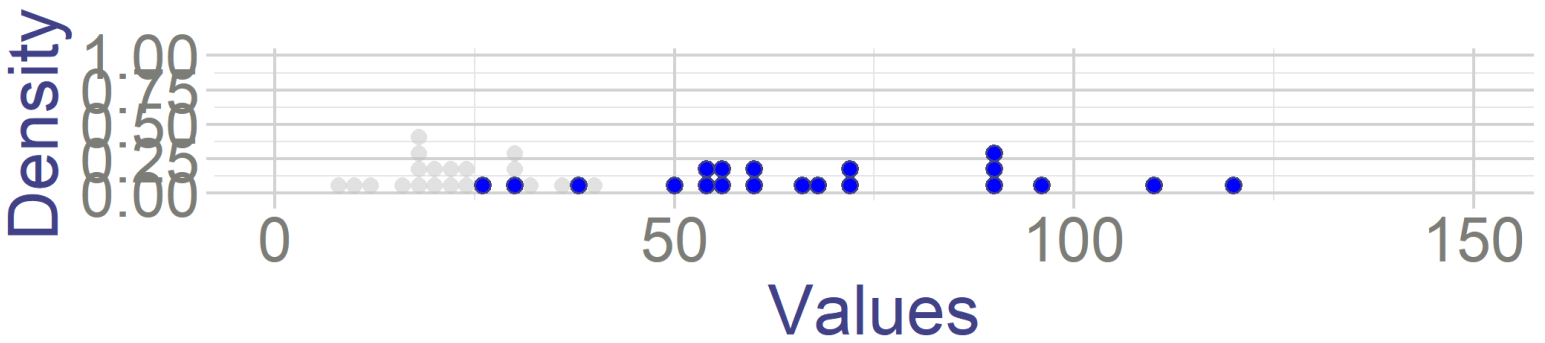
$$\mu_Y = E(Y) = \frac{\sum_{i=1}^{N}(y_i + 40)}{N} = \frac{\sum_{i=1}^{N}(x_i + 40)}{N} = \frac{\sum_{i=1}^{N} x_i + N * 40}{N} = E(X) + 40 = \mu_X + 40$$

$$\sigma_Y = \sqrt{Var(Y)} = \sqrt{\frac{\sum_i (y_i - \mu_y)^2}{N}} = \sqrt{\frac{\sum_i ((x_i + 40) - (\mu_x + 40))^2}{N}} = \sqrt{\frac{\sum_i (x_i - \mu_x)^2}{N}} = \sqrt{Var(X)} = \sigma_X$$

$$E(X + c) = E(X) + c \quad \text{and} \quad Var(X + c) = Var(X)$$

Standard Deviation

- What happens to mean and standard deviation if everyone followers get multiplied by 3? (without addition)
- $Y = 3 * X$



$$\mu_Y = E(Y) = \frac{\sum_i y_i}{N} = \frac{\sum_i 3x_i}{N} = 3\frac{\sum_i x_i}{N} = 3E(X) = 3\mu_X$$

$$\sigma_Y = \sqrt{Var(Y)} = \sqrt{\frac{\sum_i (y_i - \mu_y)^2}{N}} = \sqrt{\frac{\sum_i (3x_i - 3\mu_x)^2}{N}} = \sqrt{3^2\frac{\sum_i (x_i - \mu_x)^2}{N}} = \sqrt{3^2 Var(X)} = 3\sigma_X$$

$$E(cX) = cE(X) \quad \text{and} \quad Var(cX) = c^2 Var(X)$$

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

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

Percentiles

7. [**5 puntos**] A Mexican manufacturing company exports products to America. The monthly export revenue is a random variable with **mean** μ and **standard deviation** σ. Due to new tariffs imposed by the Amercian Government, the company's export revenues are now **halved** (i.e., each export revenue value is multiplied by 0.5). The company wants to understand how this change affects the variability of its export revenues relative to the average export revenue. Given that the export revenues are now **halved**, which of the following statements is **TRUE** regarding the coefficient of variation of the export revenues after the tariffs are imposed?

   a) The coefficient of variation will decrease by half.

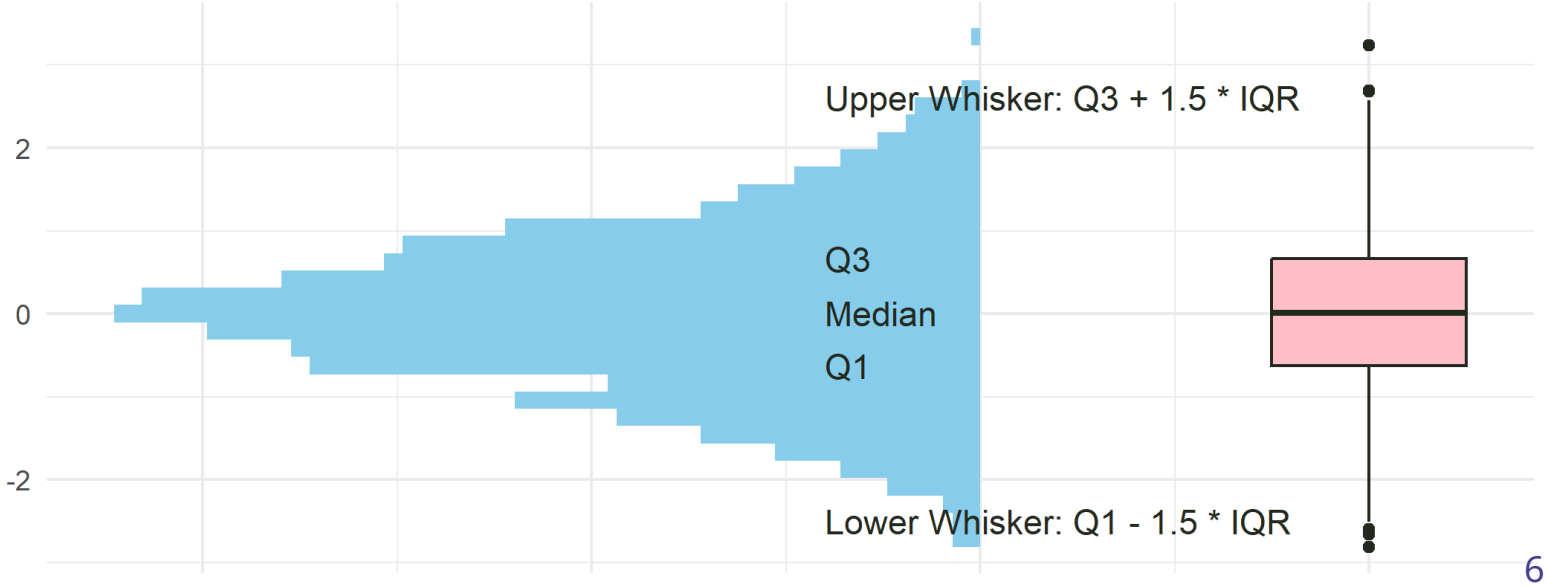
   b) The coefficient of variation will be twice as high.

   c) The coefficient of variation will **remain the same** after the revenues are halved.

   d) The coefficient of variation will **change**, but the exact change cannot be determined without recalculating.

Box and Whiskers plot

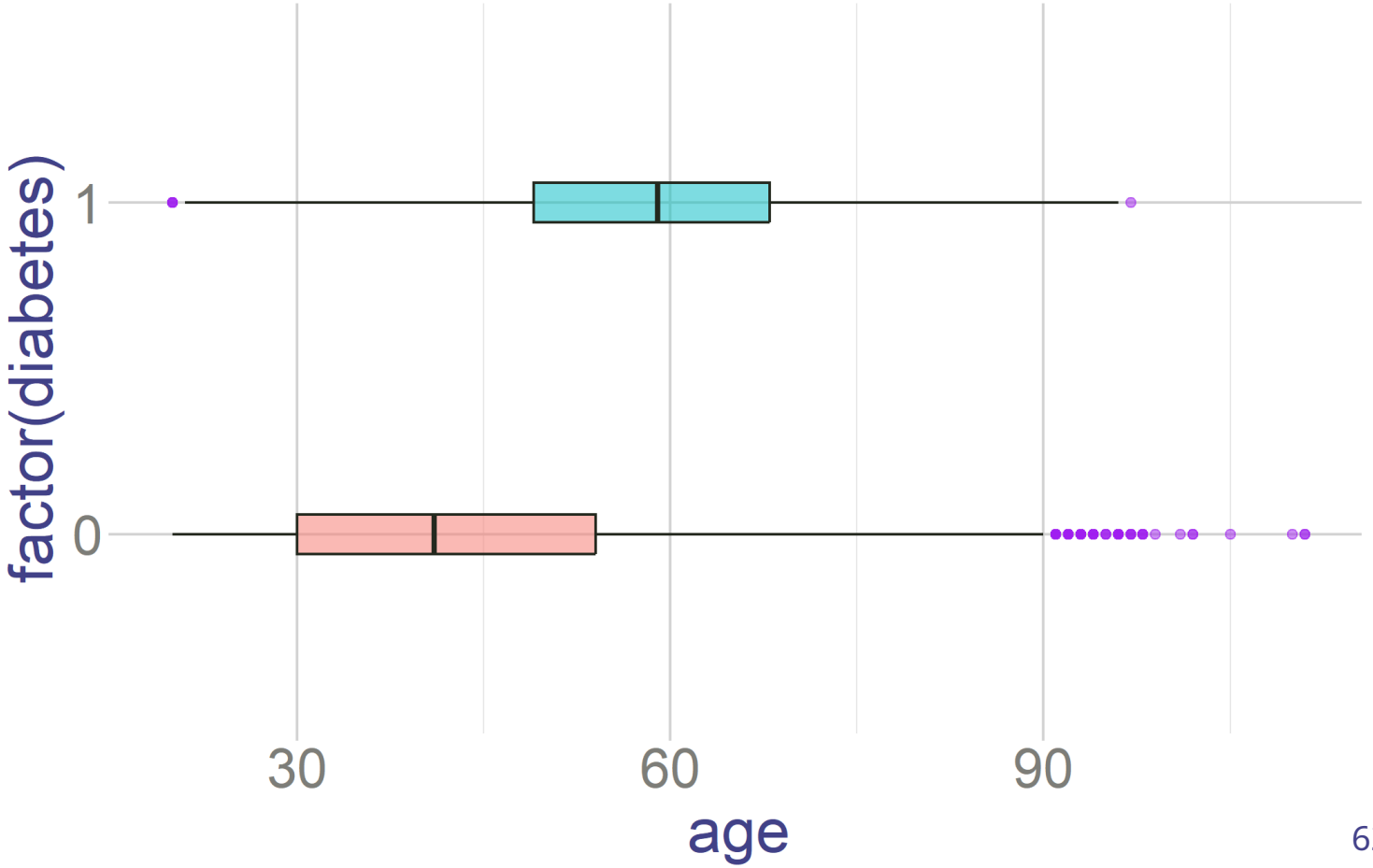
- Helps to see the distribution of the data
- Helps to see to see the outliers
  - Outliers are useful to see anomalies and potential errors in data collection
  - Whisker can be maximally 1.5 times the interquartile range
  - Any point beyond that is an outlier
  - If no point beyond 1.5 times the interquartile range, whisker goes just to the last datapoint and is shorter than 1.5 times the interquartile range



Box and Whiskers plot

Dataset comparisons

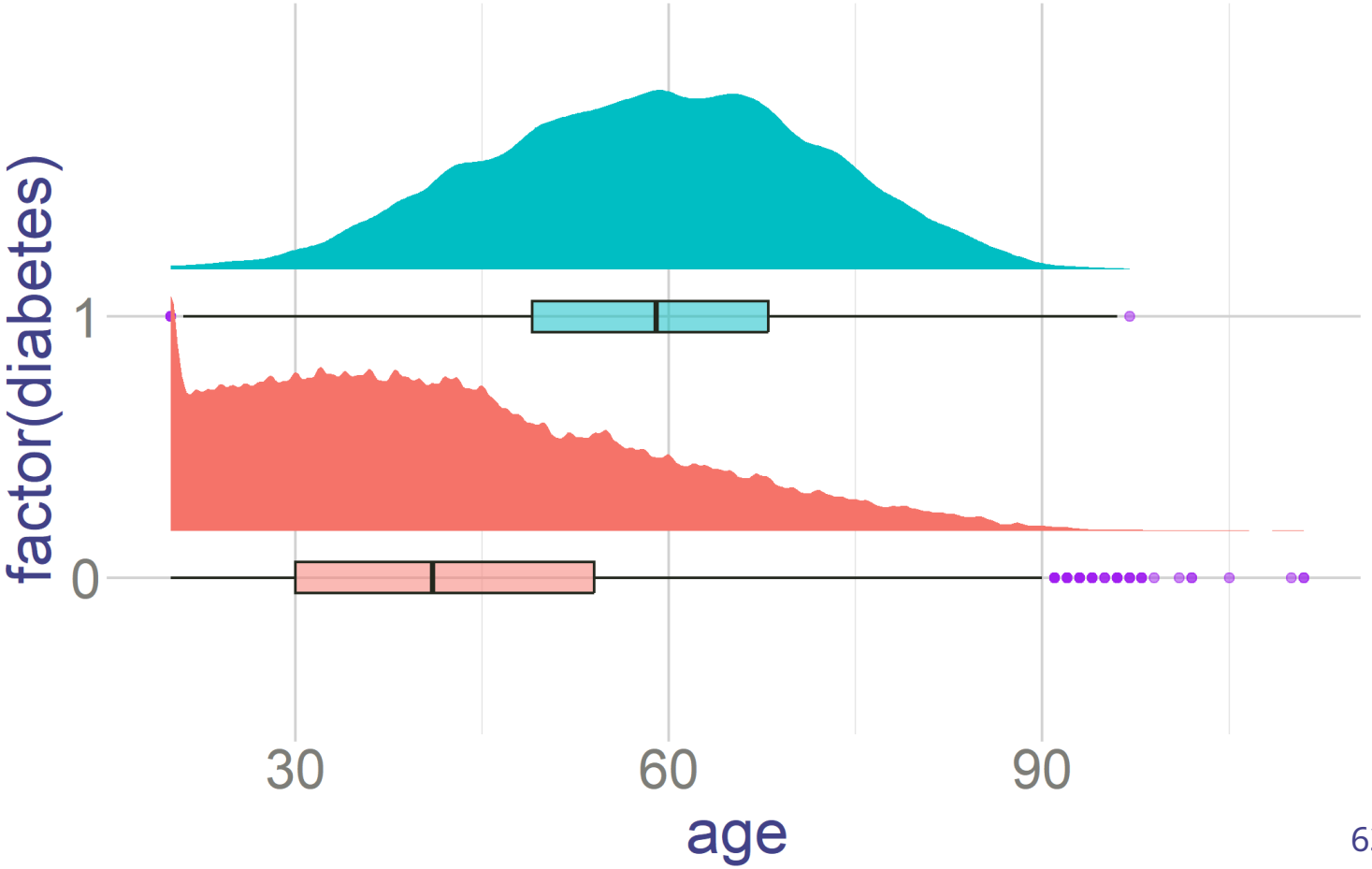
- They summarize data very well



Box and Whiskers plot

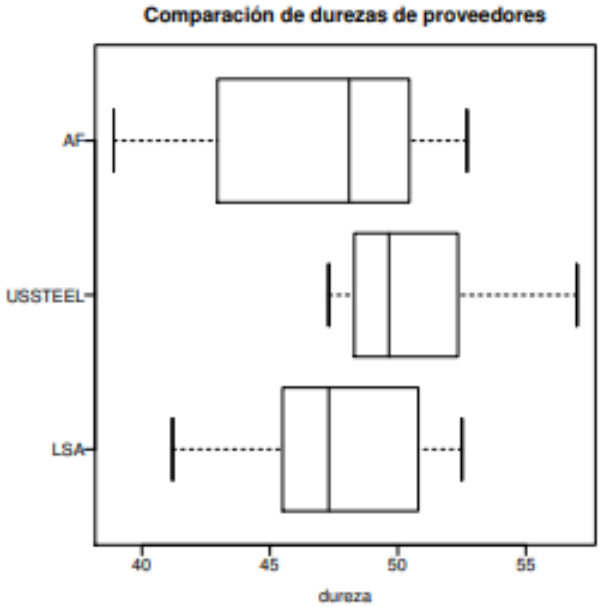
Dataset comparisons

- They summarize data very well



5. [**5 puntos**] The next plot represents data on the hardness of steel rods for three different suppliers: AF, USSteel, and LSA. Based on this figure, it may be said that:

Hardness of steel rods for three different suppliers



a) Distributions for all supplier's hardness are skewed to the left.

b) AF's hardness seems to have the least dispersion of the three.

c) AF's hardness seems to have less dispersion than that of LSA.

d) USSTEEL hardness distribution is skewed to the right.

6. [**5 puntos**] The prior plot representing the hardness of steel rods for three different suppliers, allows us to identify useful:

a) central tendency and location measures.

b) insights about the distribution's dispersion.

c) asymmetry among distributions.

Exercises:

- Review Exercises:
  - PDF 2: 1,2,6,8 (skip f),9,10,13,
- Homeworks
  - Lista 00.1: 1,2,4,5