Class 2a: Review of concepts in Probability and Statistics

Business Forecasting

Summary

- In the last class:
  - We discussed the organization of the course
  - We overviewed forecasting methods
  - We learned about methods of qualitative forecasting
  - *Reference:* Forecasting Methods and Applications, chapter 1

- This set of classes:
  - We will start learning about exploratory analysis preparing the forecast
  - We will learn to translate business problems to statistical problems
  - We will learn about various **data types**
  - We will learn how to **summarize data and observed patterns**
  - *Reference:* Forecasting Methods and Applications, chapter 2.1-2.4

Why do we care?

- To use scientific methods we need to set business problem as a statistical problem
- Define thins that we want to find to identify appropriate methods

Specific Scenario

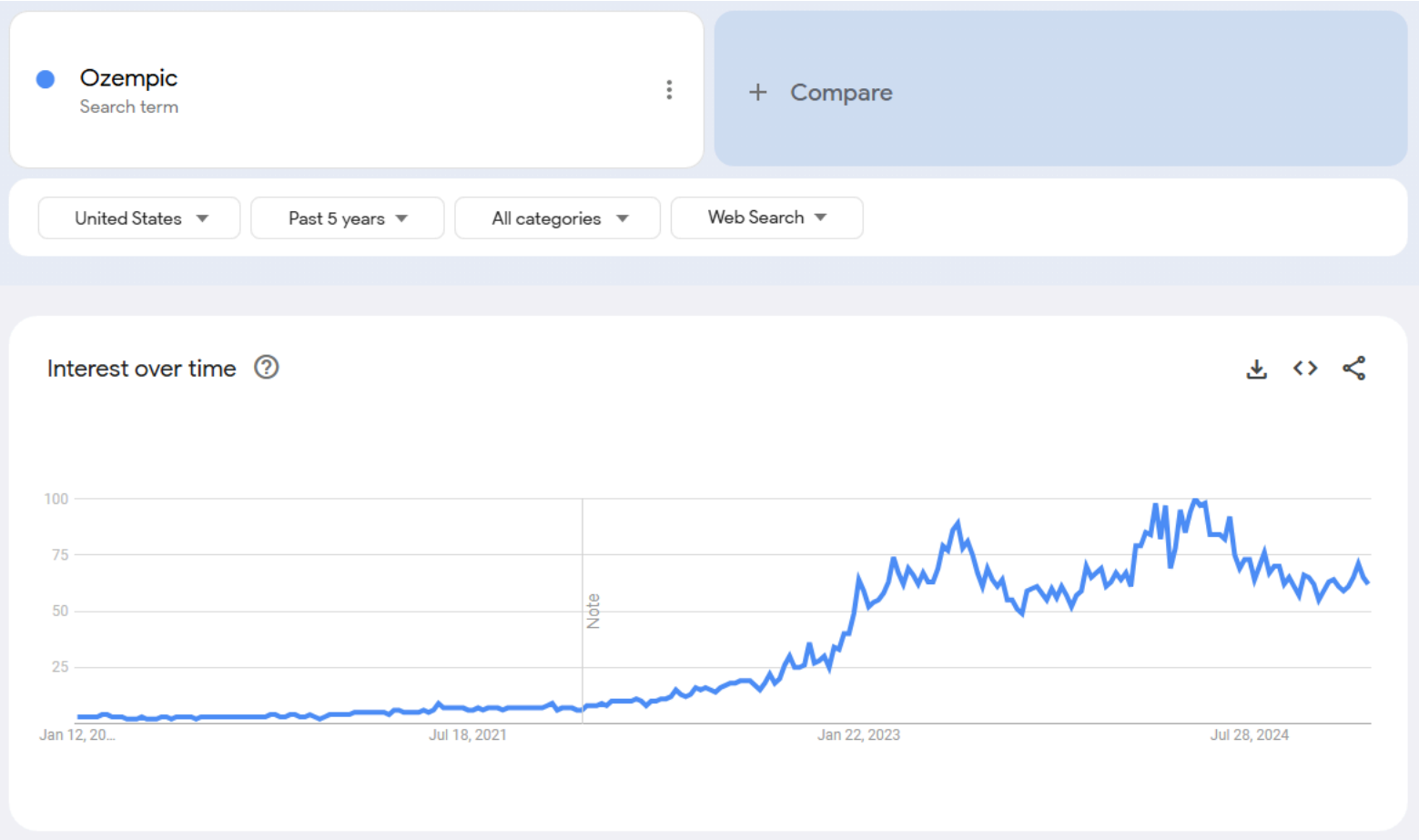
The Opportunity

- Online pharmacies are rapidly growing, offering convenience and accessibility to customers.
  - Example in Mexico: *Choiz*, which provides prescription services and drug subscriptions.

The Trend

- A new generation of highly effective anti-diabetes medications, like *Ozempic*, is gaining popularity for dual benefits: managing diabetes and aiding weight loss.

Search Interest in Ozempic



Your Role

- You have been hired as a consultant for a start-up aiming to launch a subscription service for these medications in Mexico.

The Challenge

- Your boss needs a detailed exploratory market analysis to forecast sales and identify key customer segments

Parameters vs Statistics

- You need to know how many people in Mexico have diabetes

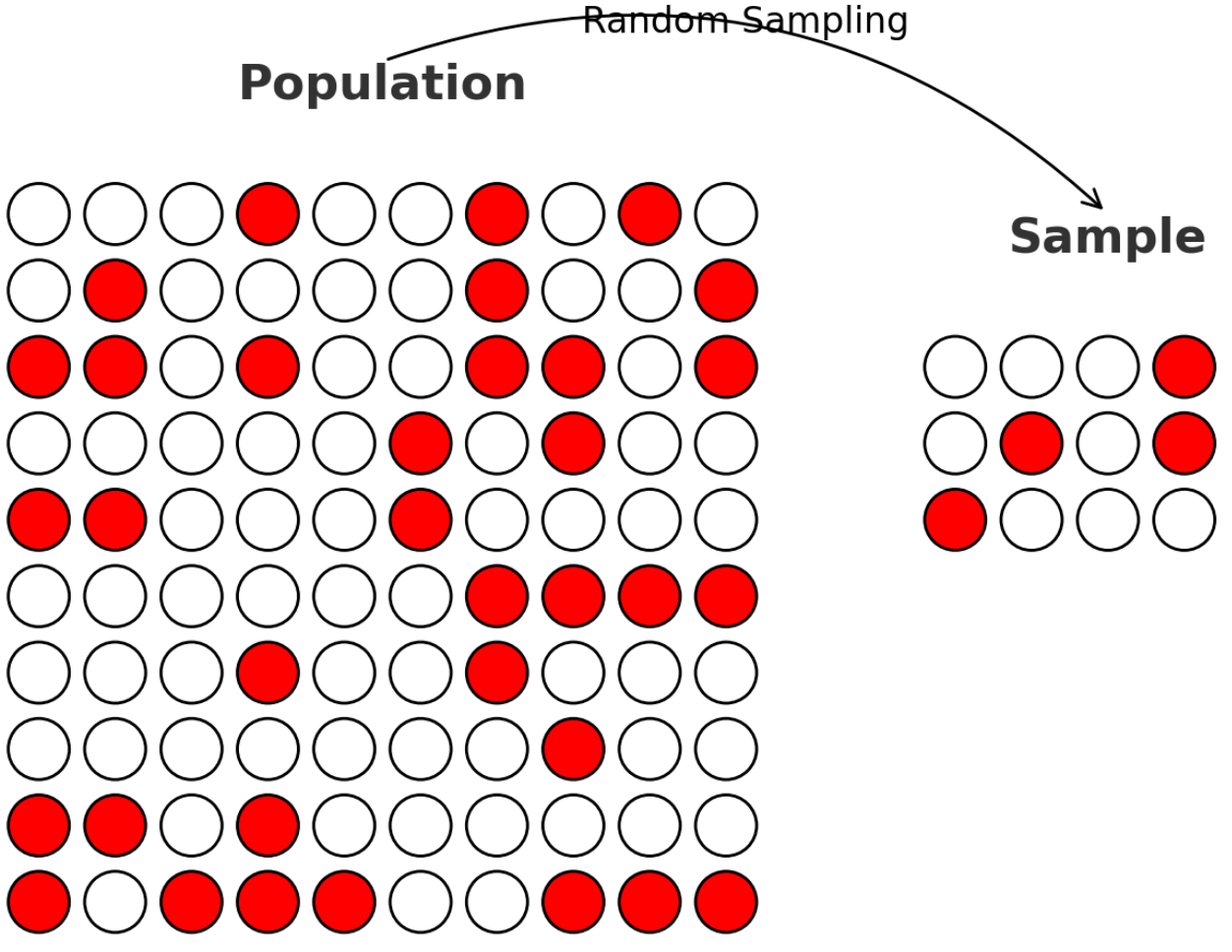
  Parameter

- Call μ_d the proportion of Mexican population which has diabetes
  - Usually the parameter is an **unknown** number **describing the whole population**
  - You want to learn what it is
  - In our example, μ_d is a parameter that you want to learn
  - More generally, parameter describes an aspect of the entire population

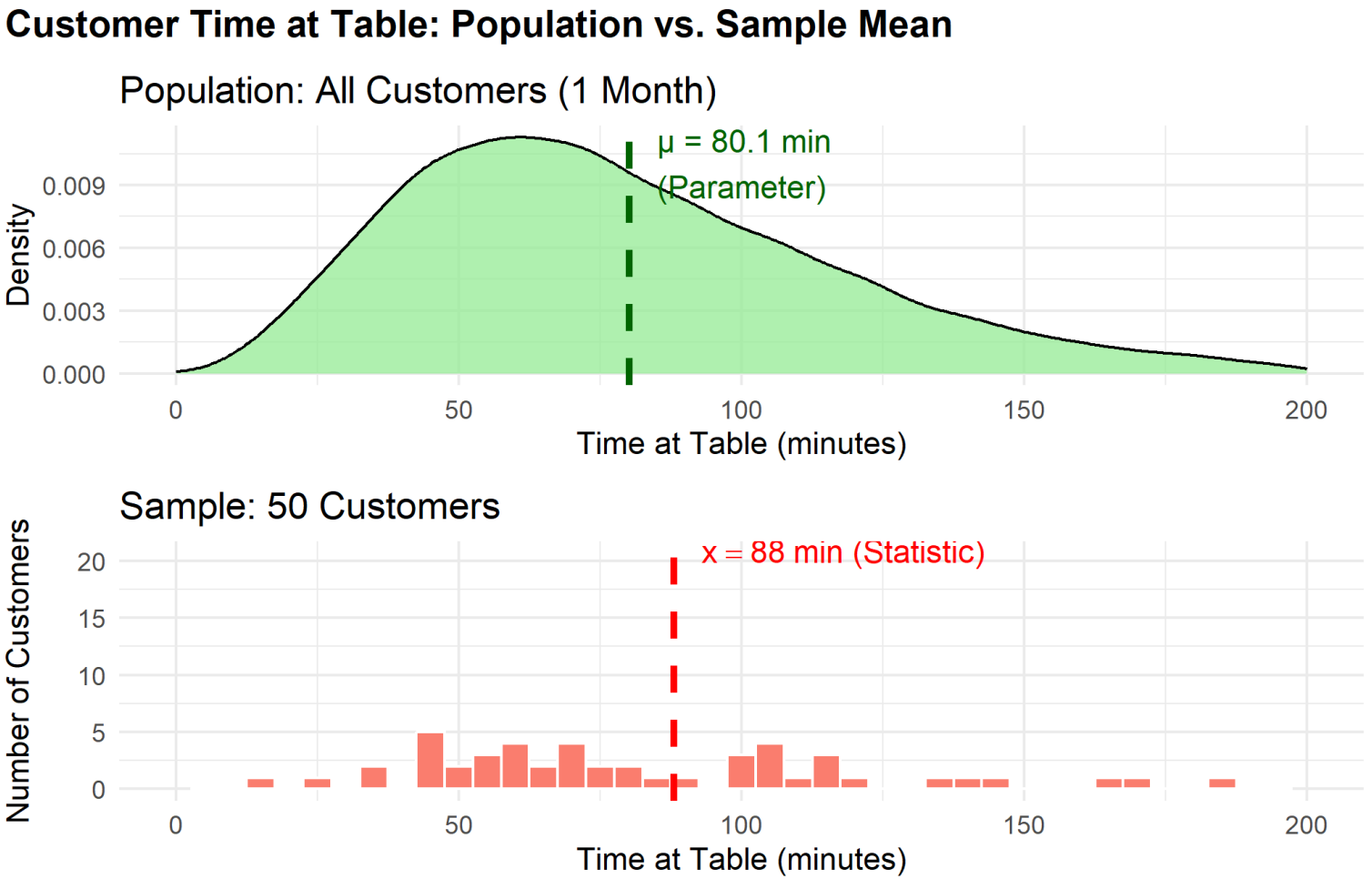
  Statistic

- But you don't have data on the whole population. At best you can get a sample from a survey
  - So you will try to estimate this parameter with sample
  - Statistic is a **guess of the parameter** which can be **calculated from the sample**
  - You will calculate a statistic $\hat{\mu}_p$ which is the proportion of diabetics in the sample

Sampling: categorical variable



Sampling: continuous variable



Parameters vs Statistics

- What is population, sample, parameter and statistic in the following examples?
- You want to know the probability that a user who got a match on tinder will go out on a date with that person. You survey 1000 users and ask them about each match they got if they went on a date. You then calculate the share of dates which ended up in a match for these users.
- You want to know what whether starbucks baristas are faster than Cielito Querido baristas. You go to 10 starbucks and 10 Cielito Querido and measure the time it takes to make a coffee. You then calculate the average time it takes to make a coffee in each of these chains.
- You want to know the average age of people who go to the gym. You go to a gym and ask 100 people about their age. You then calculate the average age of these people.
- You want to know the variance of internet speed during in Mexico City. You visit 500 households and calculate the variance of their internet speed.

Parameters vs Statistics

- What are some examples of the statistics we will use and why do we care?

Summary Statistics

Summary Statistics

- Any advanced analysis always starts with describing the data

    1. **Cambridge Analytica**

- Analyzed voter demographics and turnout patterns to target swing voters with tailored political ads.

    1. **Tesla**

- Summarized battery efficiency and charging times to improve EV performance and user experience.

    1. **Walmart**

- Optimized opening hours and staffing by analyzing distribution

    1. **BBVA**
- Analyzed average credit card balances and repayment trends to design tailored credit offers and improve customer retention.

Basically any modern company relies on summarizing data to make decisions

Measures of Central Tendency

Mean

- **Mean** represents the arithmetic average of the data.
- Sometimes called the expected value of the random variable E(X)
- The population mean μ is the sum of all observations divided by the total population size:

$$\mu = E(X) = \frac{\sum_{i=1}^{N} x_i}{N} = \sum_{x \in X} P(X = x) \times x$$

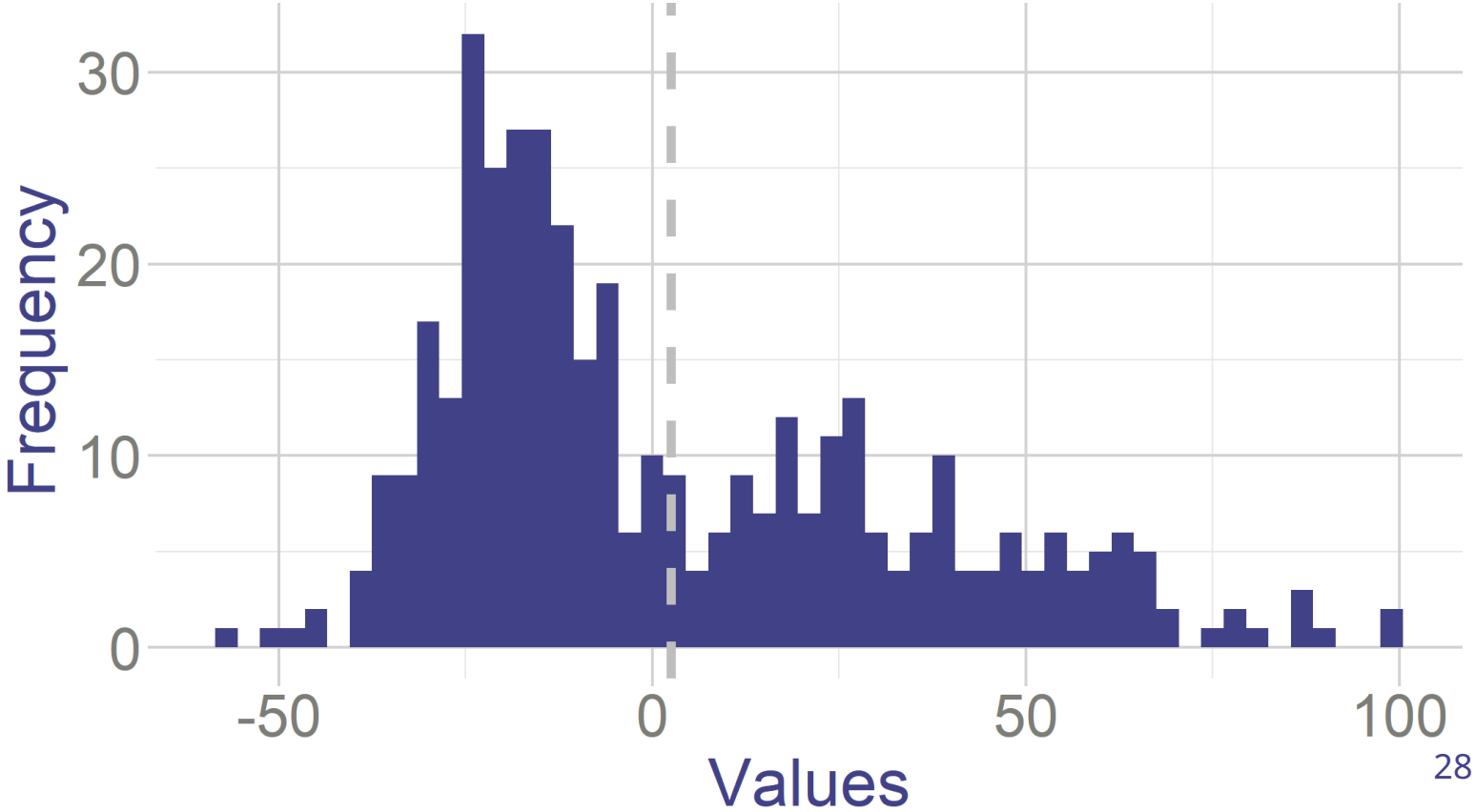
- where N is the total population size, and x_i are individual data points.
- The sample mean, denoted as $\bar{x}$, is the sample equivalent:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + \ldots x_{n-1} + x_n}{n}$$

where n is the sample size.

Mean

Intuitively, mean is the balancing point of the distribution.



Mean of a binary variable

What if a mean of a **binary variable**?

- Binary variable is a variable which takes value 0 or 1
- For example: do you have diabetes (yes=1, no=0)

What is the intuitive interpretation of the mean of this variable?

- $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
- $\bar{x} = \frac{1+0+0+...0+1}{n} = \frac{n_{diabetes}}{n} = \hat{\mu}_{diabetes}$

It's the proportion of people with diabetes in the sample: mean(diabetes)= 0.11

Weighted Mean

- In some scenarios, data points have different weights.
- For a dataset with weights w_i and values x_i, the weighted mean is:

$$\text{Weighted Mean} = \frac{\sum_{i=1}^{n} w_i \cdot x_i}{\sum_{i=1}^{n} w_i}$$

Show 4 entries

Person	Weight	Grade
Midterm 1	0.2	6
Midterm 2	0.2	8
Quizzes	0.15	9
Final Project	0.15	4

Showing 1 to 4 of 5 entries

Previous 1 2 Next

The **weighted mean** is:

$$\bar{x} = \frac{0.2 \times 6 + 0.2 \times 8 + 0.15 \times 9 + 0.15 \times 4 + 0.3 \times 8}{0.2 + 0.2 + 0.15 + 0.15 + 0.3}$$

Mean

- Is mean always a right measure?

"Bill Gates walks into a bar"

- Suppose a group of people, including Bill Gates, walks into a bar.
- Let's say the net worth of everyone in the group is as follows:

Show 6 entries

Person	Net_Worth
Person 1	10
Person 2	20
Person 3	30
Person 4	40
Person 5	50
Bill Gates	600000

Showing 1 to 6 of 6 entries Previous 1 Next

The **mean** is:

$$\bar{x} = \frac{10 + 20 + 30 + 40 + 50 + 60000}{6}$$

$$= 10025$$

Mean is seriously skewed due to the outlier.

Mean vs Median

Median

- **Median** represents the middle value when data is sorted
- Half of observations are below it, half are above it.
- For a dataset with odd size n, the median is the $\frac{n+1}{2}$-th value
- For even size n, it's the average of $\frac{n}{2}$-th and $\frac{n}{2} + 1$-th values.

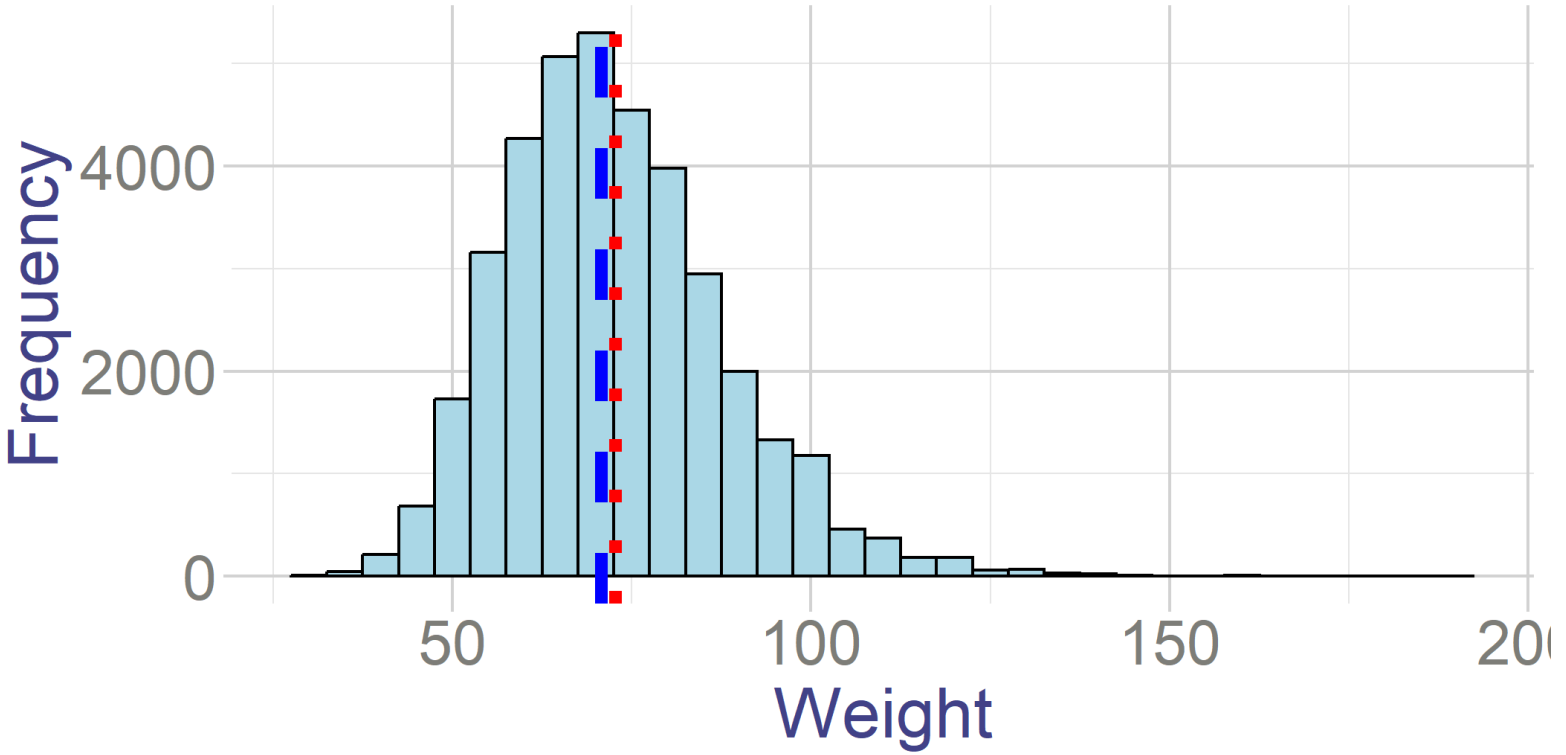
Day	Number of Customers
1	20
2	18
3	25
4	22
5	30
6	21
7	27

The dataset has $n = 7$ (odd) observations, so to find the median:

- Arrange the data in ascending order:
  - 18, 20, 21, 22, 25, 27, 30.
- The median is the $\frac{n+1}{2}$-th value, which is the 4th value.
- Thus, the median is the 4th value, which is 22.

Let's look at the median weight in our population

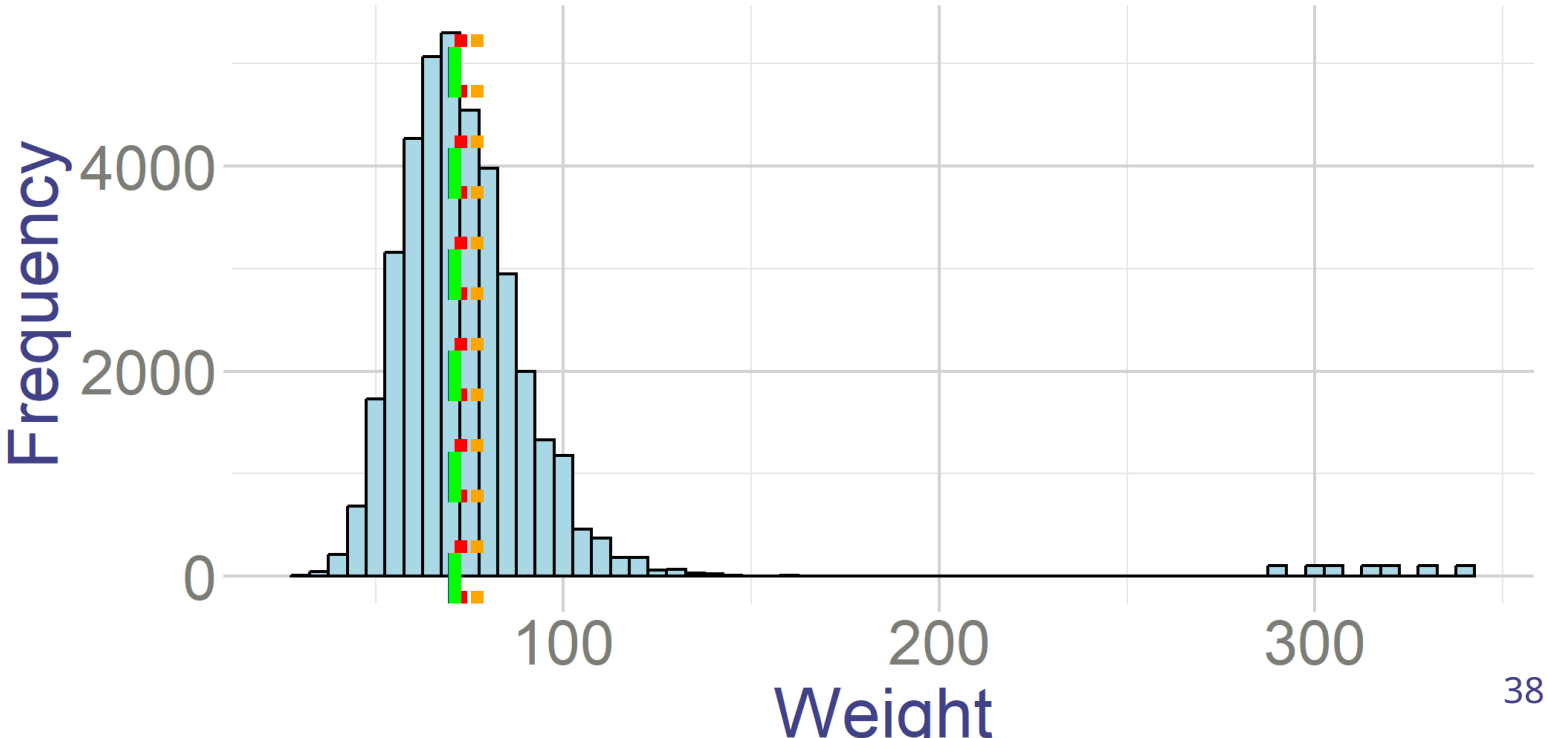
- Mean: 72.66451
- Median: 70.7536



Median and outliers

I added couple of observations on the right tail of the distribution

- Old Mean: 72.66, **New Mean: 77.05**
- Old Median: 70.75, **New Median: 70.95**



Side note on the Mode

Mode is the most frequent value in the data

- Let's look at the distribution of age of people with diabtese

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

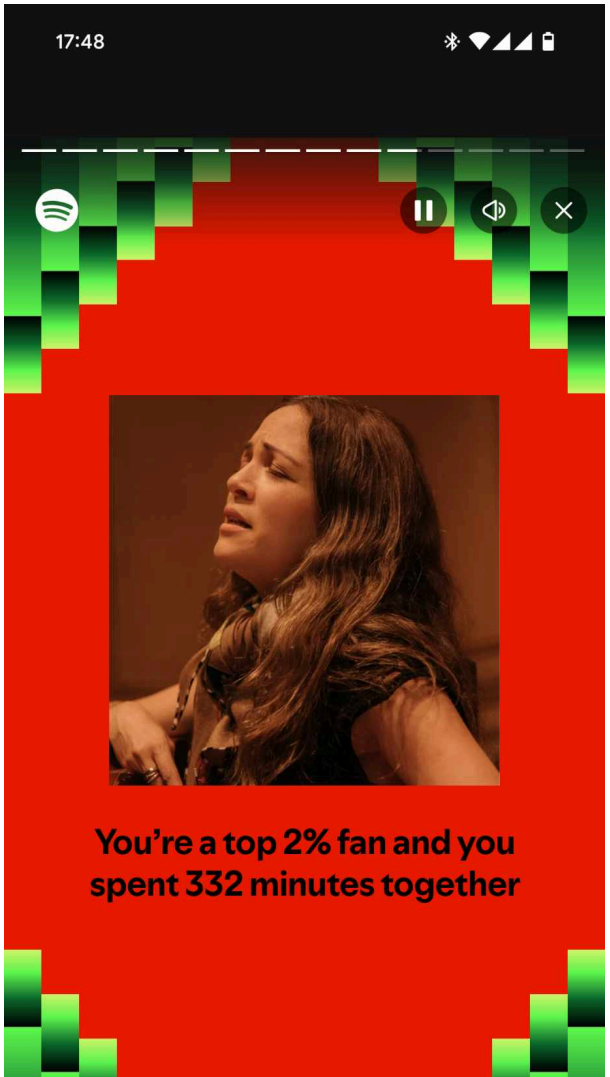
Showing 1 to 6 of 78 entries

Previous 1 2 3 4 5 ... 13 Next

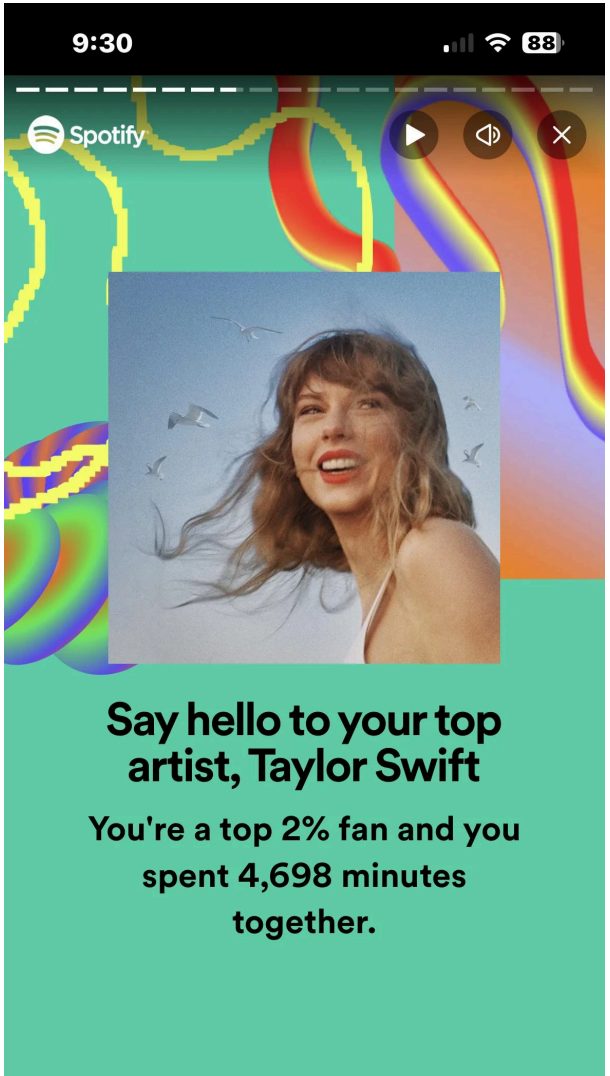
Mode


Mean
People only use me when data is normal
Median
They only use me when there are outliers
Mode
You guys are getting used

Percentiles

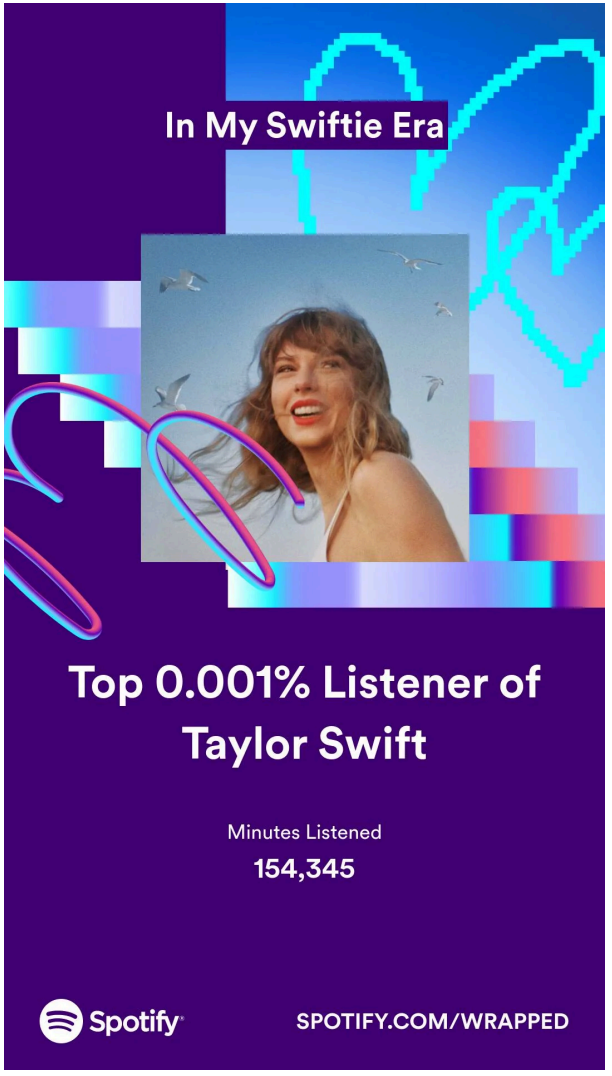


Percentiles

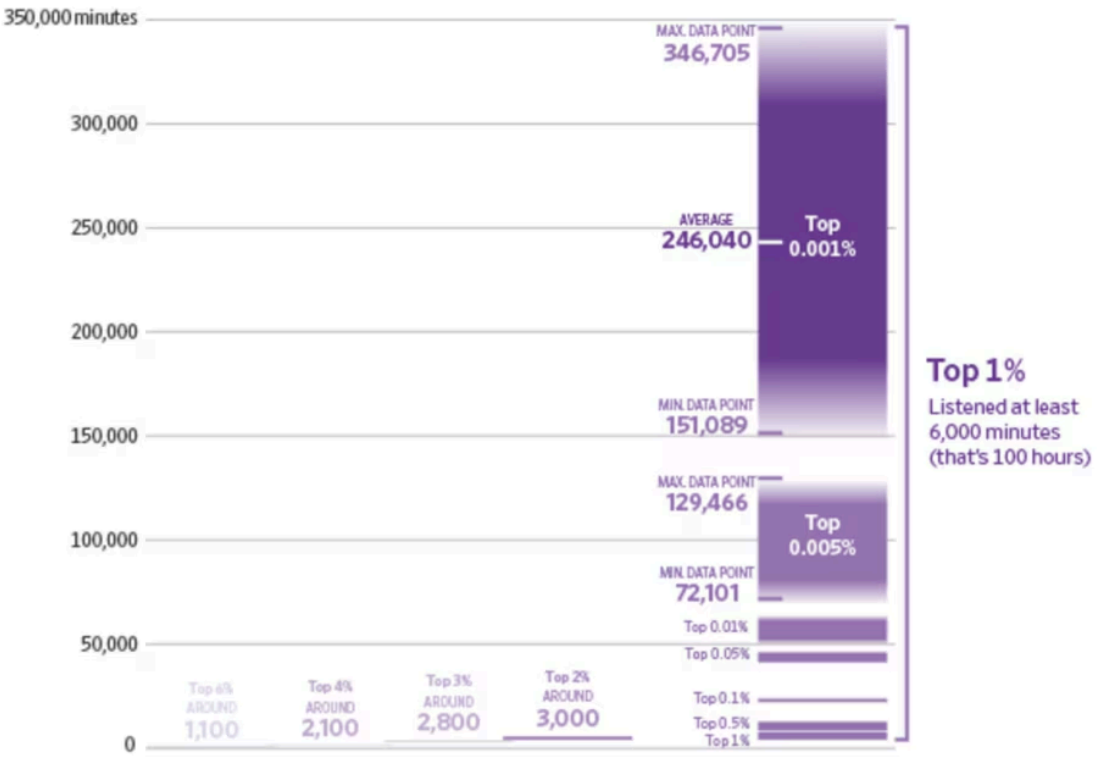


Percentiles

Percentiles



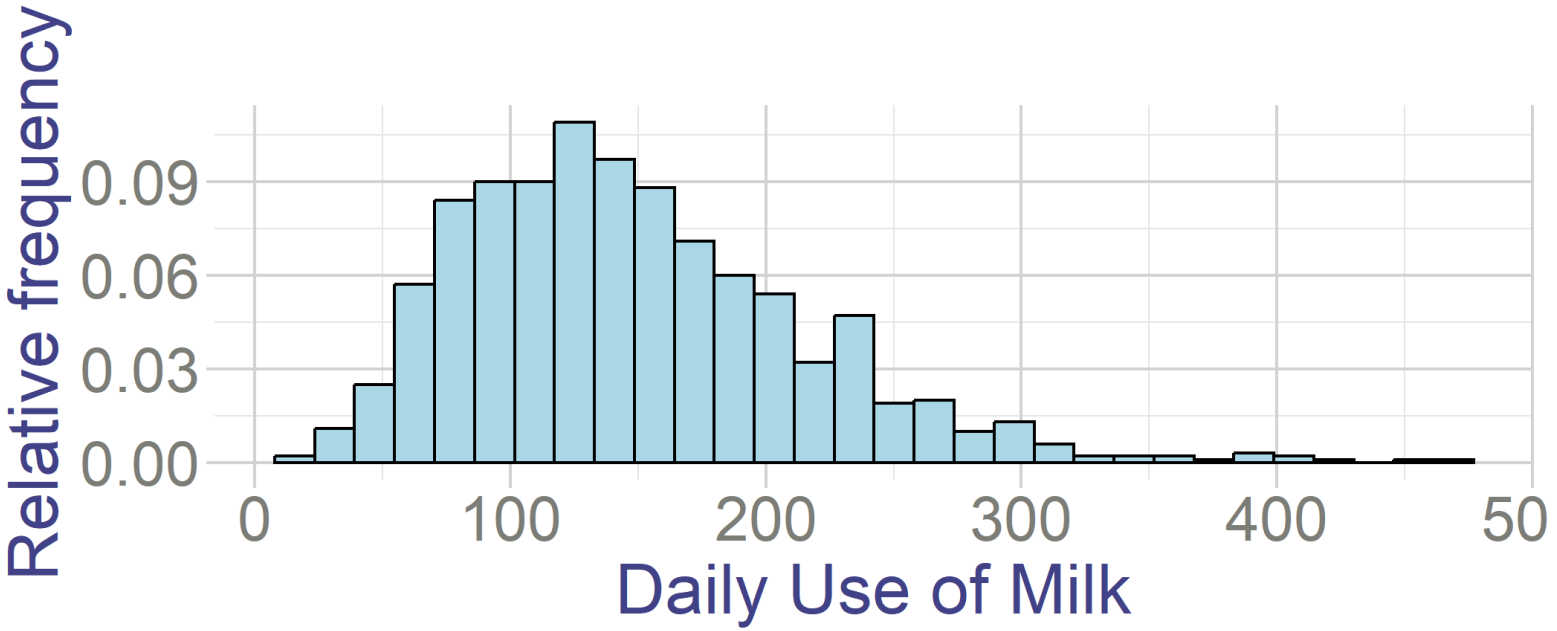
Percentiles



Credits: Wall Street Journal Article Link

Percentiles

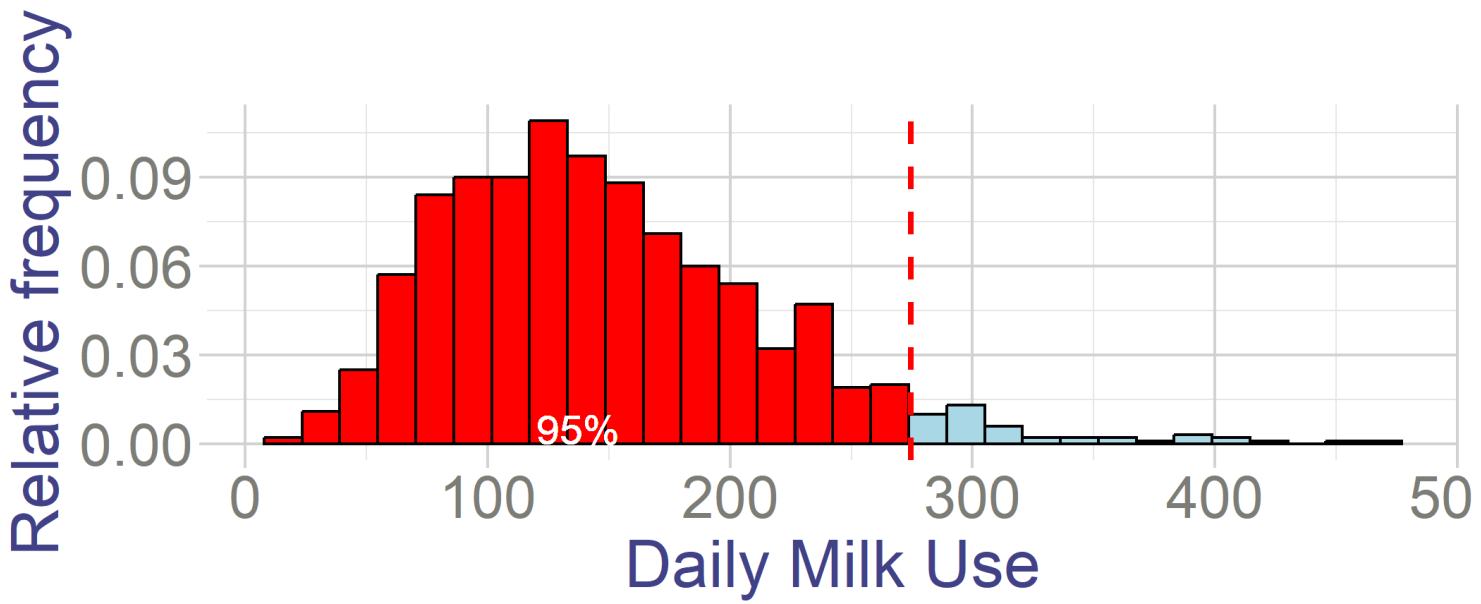
- How much inventory of milk you need to keep in your Starbucks?
- What is the tradeoff of keeping too much vs too litle inventory?
- Suppose we want to have enough of milk to cover sales on 95% of days
- To figure it out, let's look at the distribution of the daily use of milk



Percentiles

- Let s_i be the daily sales of milk
- We want to choose amount M, such that $P(s_i \leq M) = 0.95$

- That is, in 95% of days sales are smaller or equal than M



- What is this number?

- It's the 95th percentile of the distribution (274 liters)

Percentiles

- *Percentiles* divide the ordered data into 100 equal parts.
- pth percentile is a value such that $p\%$ of the data are below it
  - v_p is such that $P(x_i \leq v_p) = p$
  - v_{95} is such that $P(x_i \leq v_{95}) = 95\%$

Percentiles

- What is the the height such that 75% of ITAM students are smaller than this height?
- What is the income level such that 25% of people in Mexico earn less than that level?
- What is the age, such that 50% of people die before that age?

How to find it in a sample

1. Arrange the data in ascending order
2. Find which observation corresponds to the relevant percentile
    - Formula: $i = \left(\frac{p}{100}\right)(n+1)$
    - Example: To find 95th percentile in a sample of 1000 observations we look at $i = \left(\frac{95}{100}\right)(1000+1) = 950.95$ observation
3. If it's an integer, value of ith observation is your percentile
4. If it's not, take the average between ith rounded down and ith rounded up
    - In our example it would be the average of 950th and 951th observation

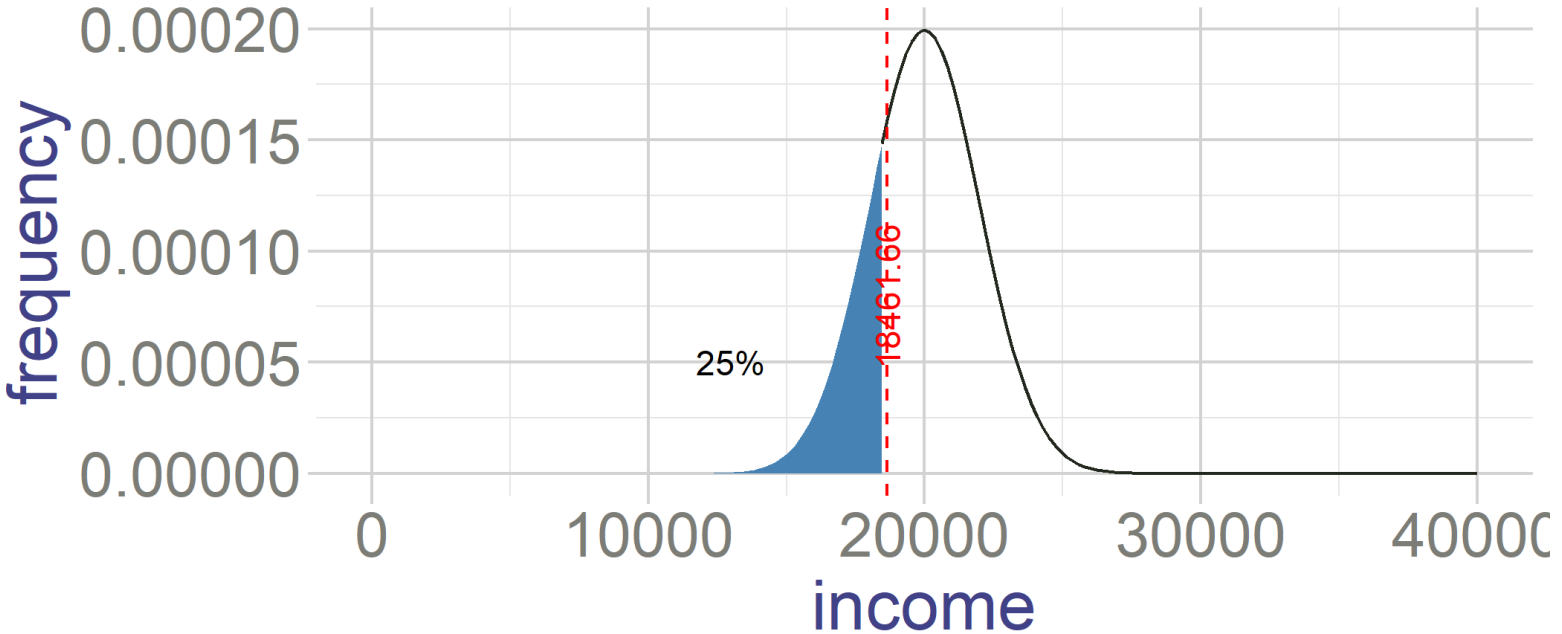
--

Or use the CDF

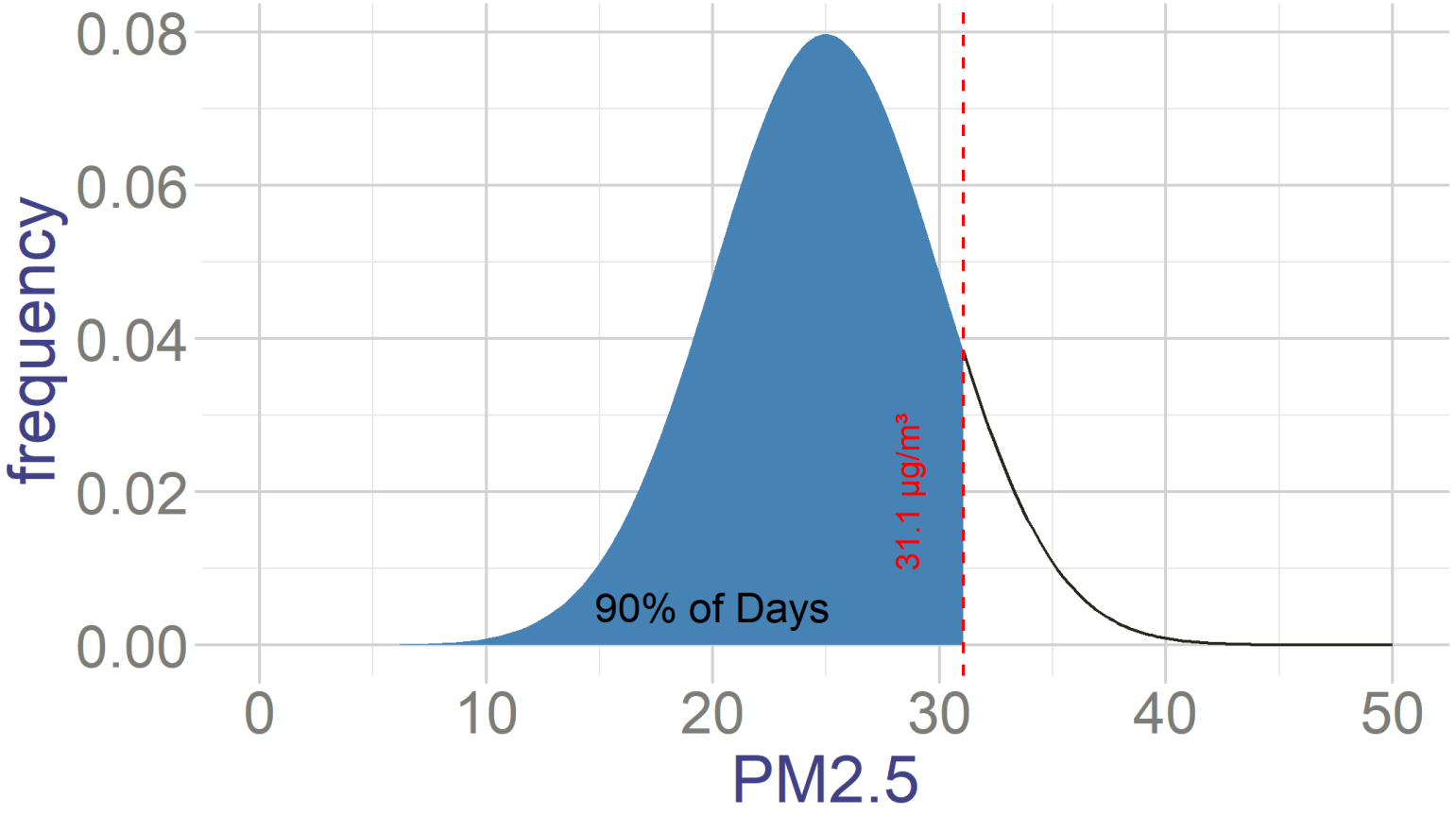
- $ECDF(v) = P(x_i \leq v)$

Common values

- **Median** - 50th percentile - half of the values are below the median
- **Quartiles** - 25th, 50th and 75th percentile.
  - How poor is the poorest quartile of the society?
  - Their income is below the 25th percentile



- **Deciles** - 10th, 20th, ... 90th
  - How bad pollution gets in CDMX during top 10% polluted days?
  - During top 10% of polluted days pollution level is larger or than 9th decile.



Example with data

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

- What is the 1st decile?
- What is the 95th percentile?

Show 4 entries

VideoTitle	Views
TikTok Video 1	172204
TikTok Video 2	9442
TikTok Video 3	37975
TikTok Video 4	56914

Showing 1 to 4 of 200 entries

Previous 1 2 3 4 5 ... 50 Next

- Index for the first decile is: $i = \left(\frac{10}{100}\right)(200+1) = 20.1$
  - First decile is the average of the 20th and 21st observation
- Index for the 95th percentile is: $i = \left(\frac{95}{100}\right)(200+1) = 190.95$
  - 95th percentile is the average of the at 190th and 191st observation

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!

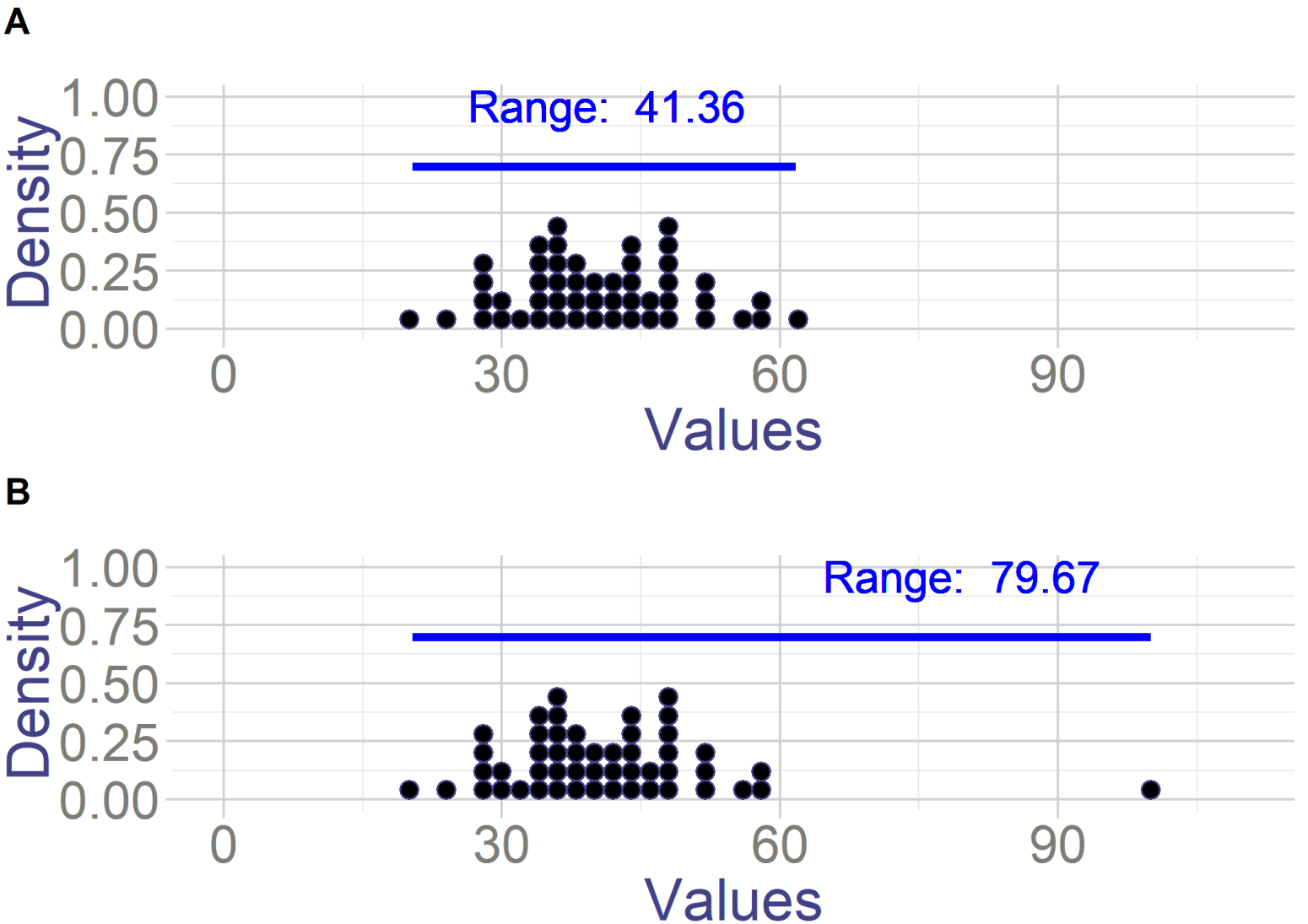
Range

- **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

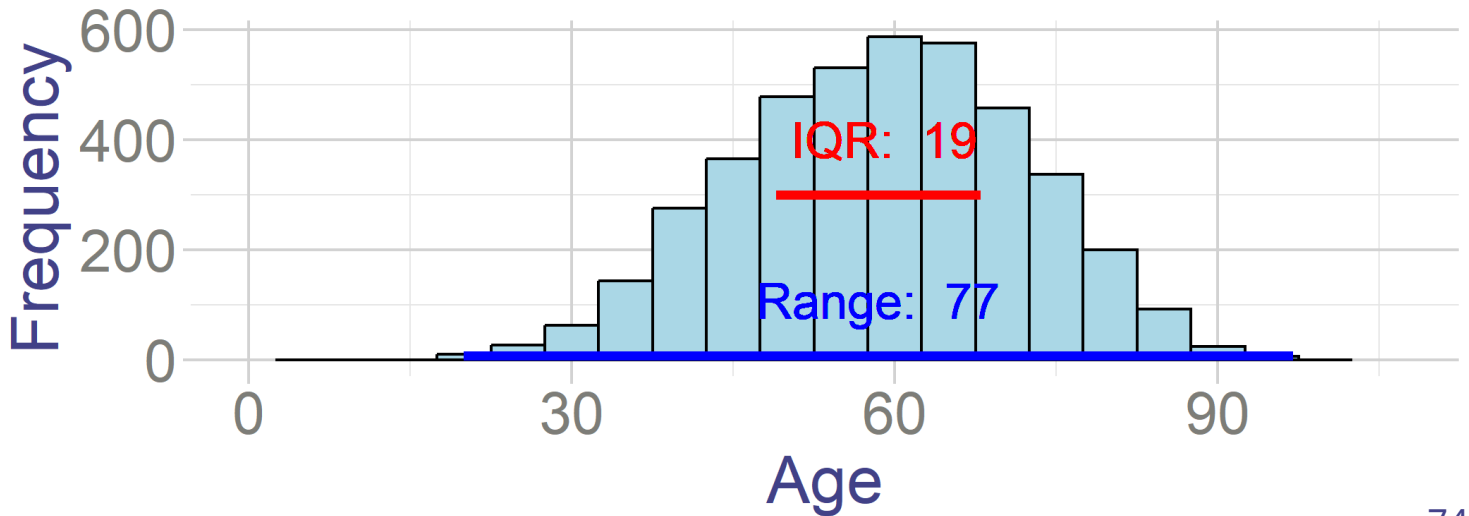


Interquartile Range

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

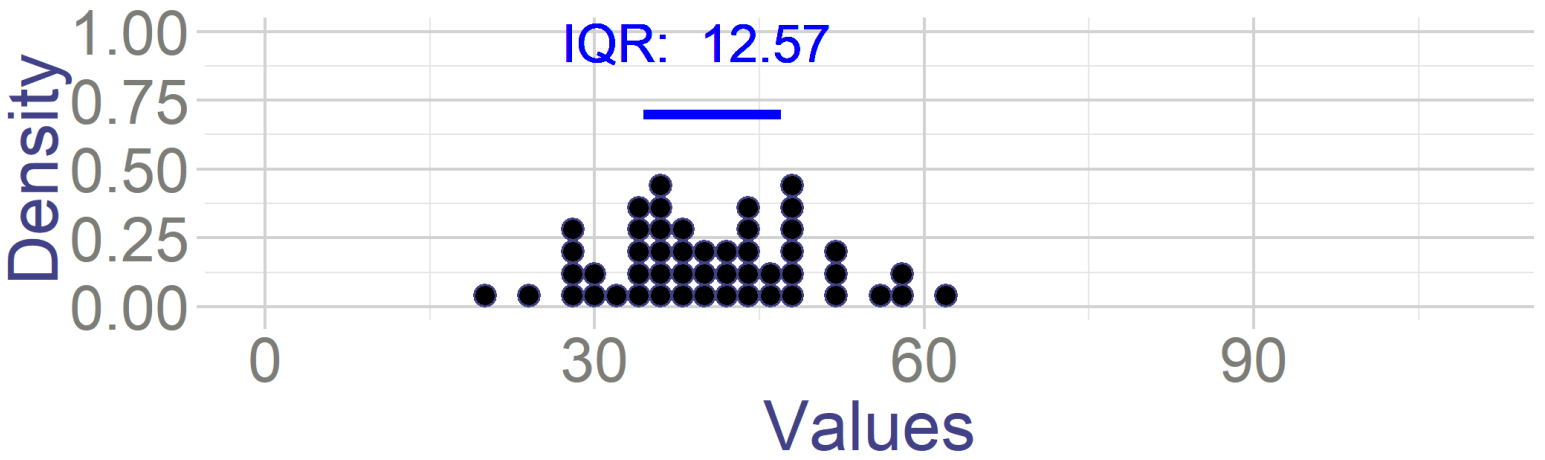
$$IQR = q_3 - q_1$$

- 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

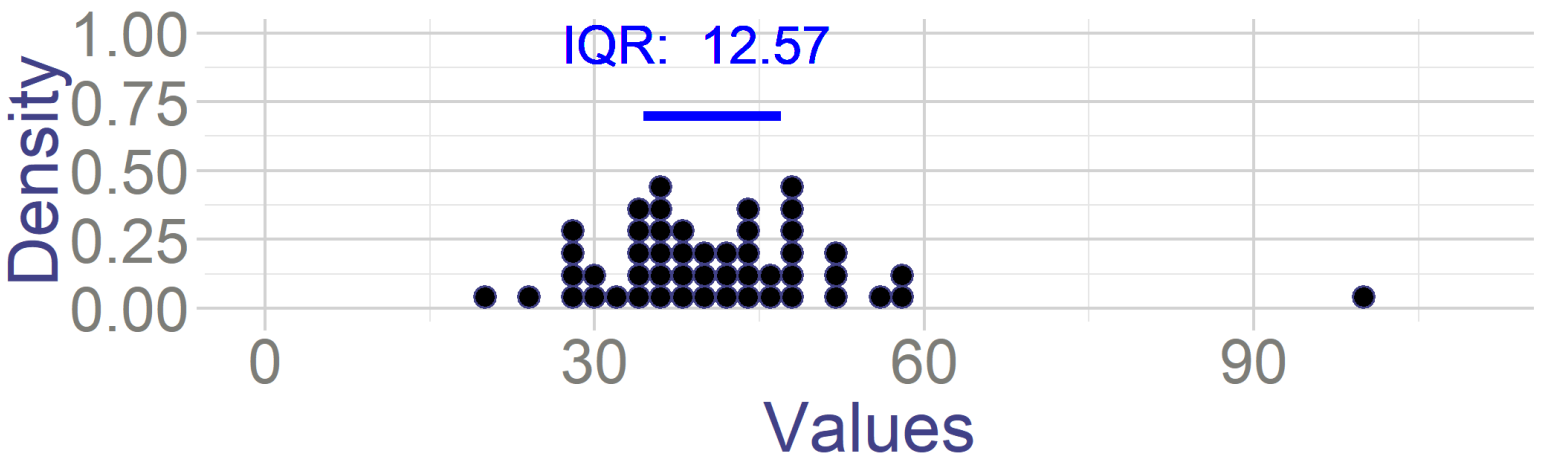


- Is it more or less sensitive to outliers than range?

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

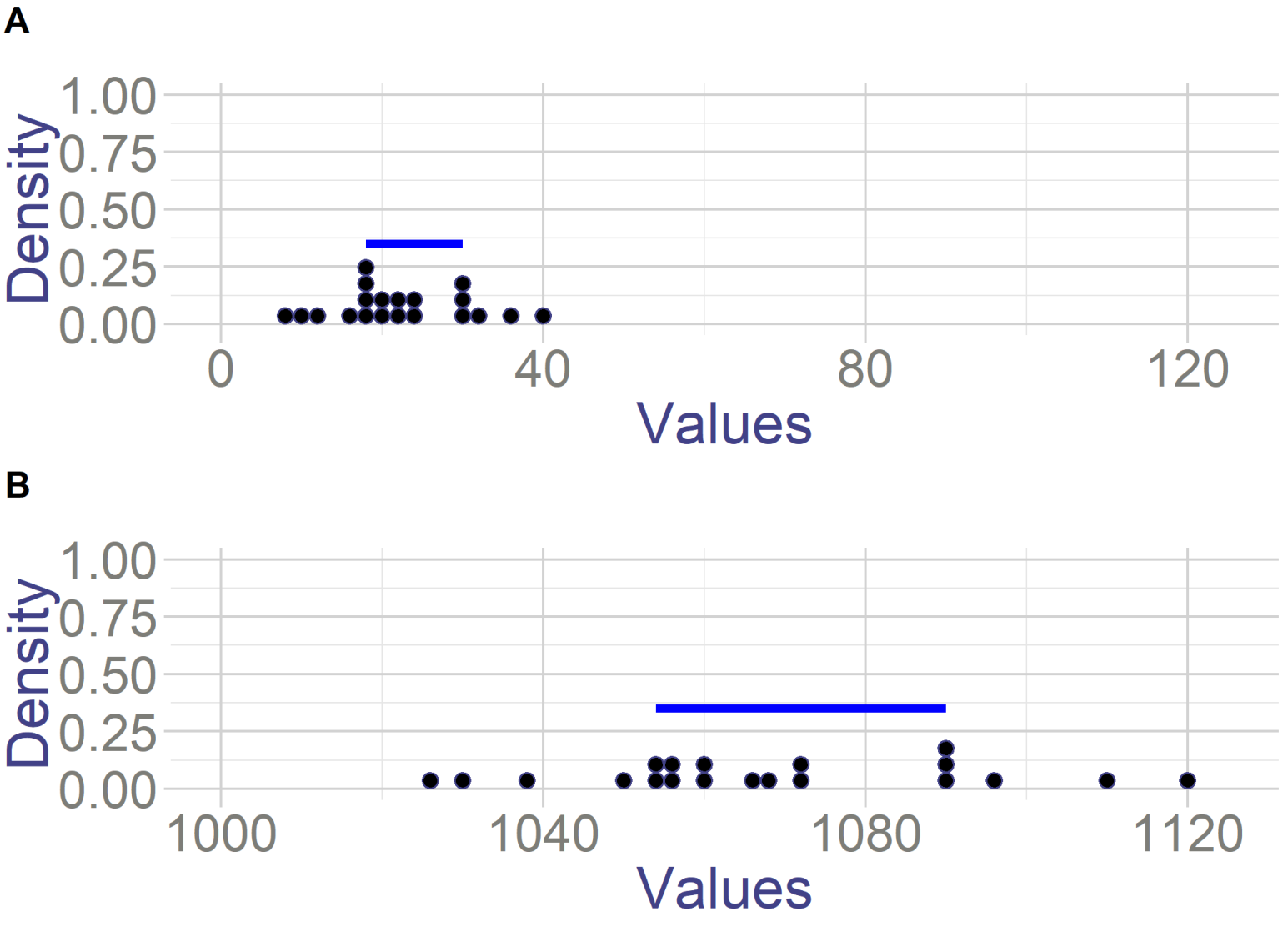
$$y_i = bx_i + a$$

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?

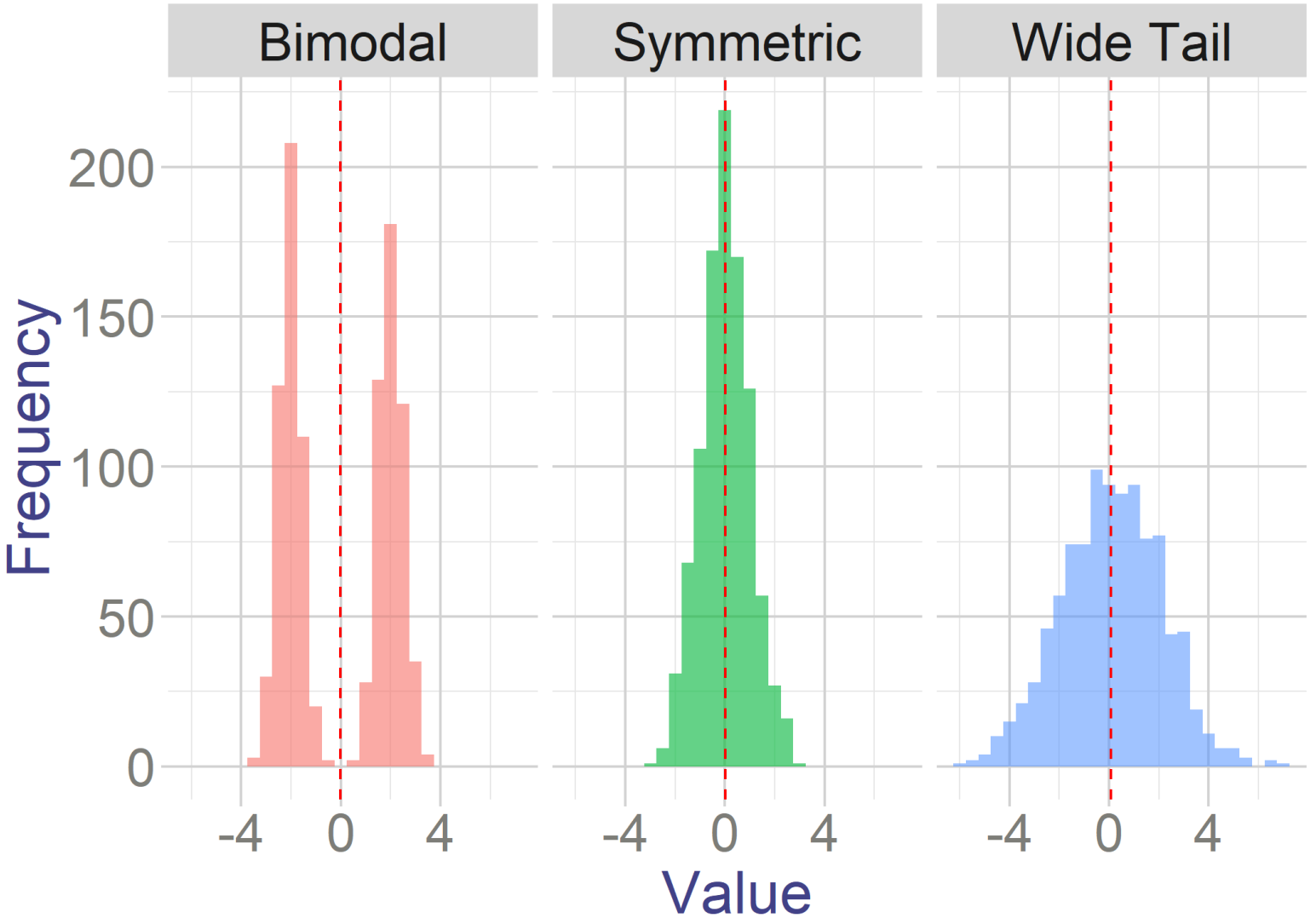
  - $MAD = frac1N \sum_{i=1}^{N} |(x_i - \mu)|$

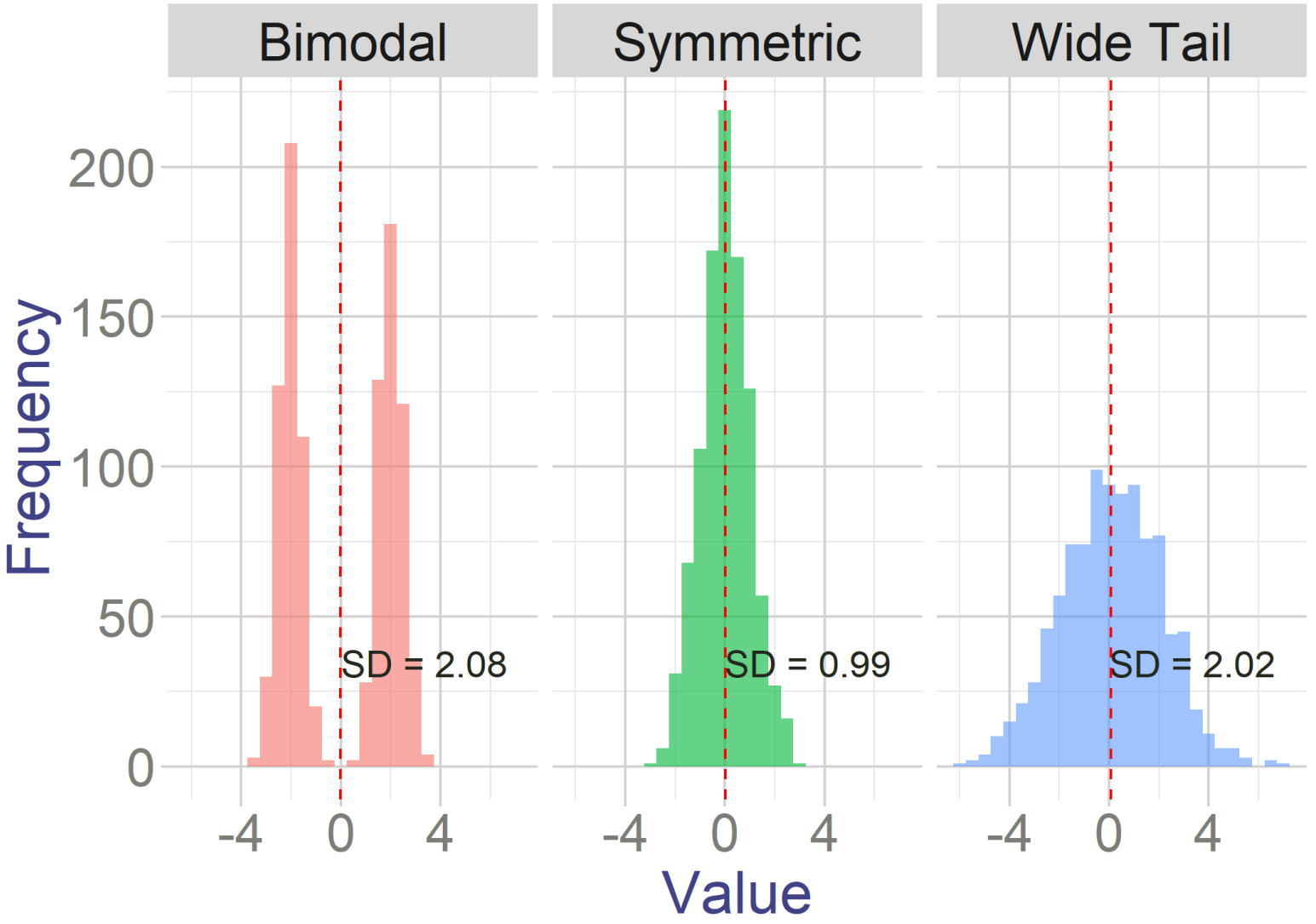
  - MAD is not differentiable at 0 :(

- Variance puts more weight on far away observations.
- It's a weighted average distance, where weights are distances themselves.

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

- It makes it very sensitive to outliers, which get a lot of weight!
- Standard deviation retains this property


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**:

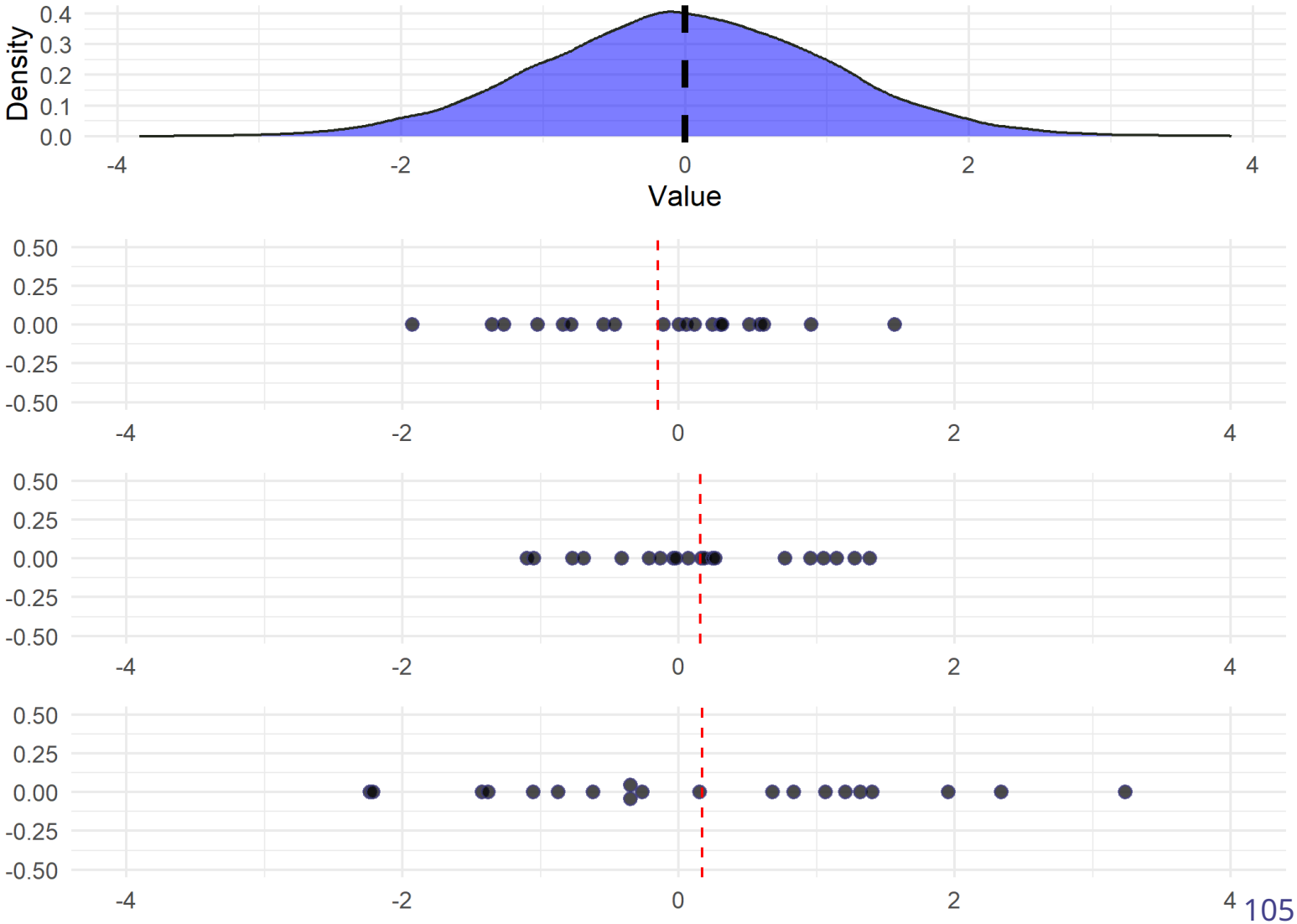
$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2)$$

  - **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?

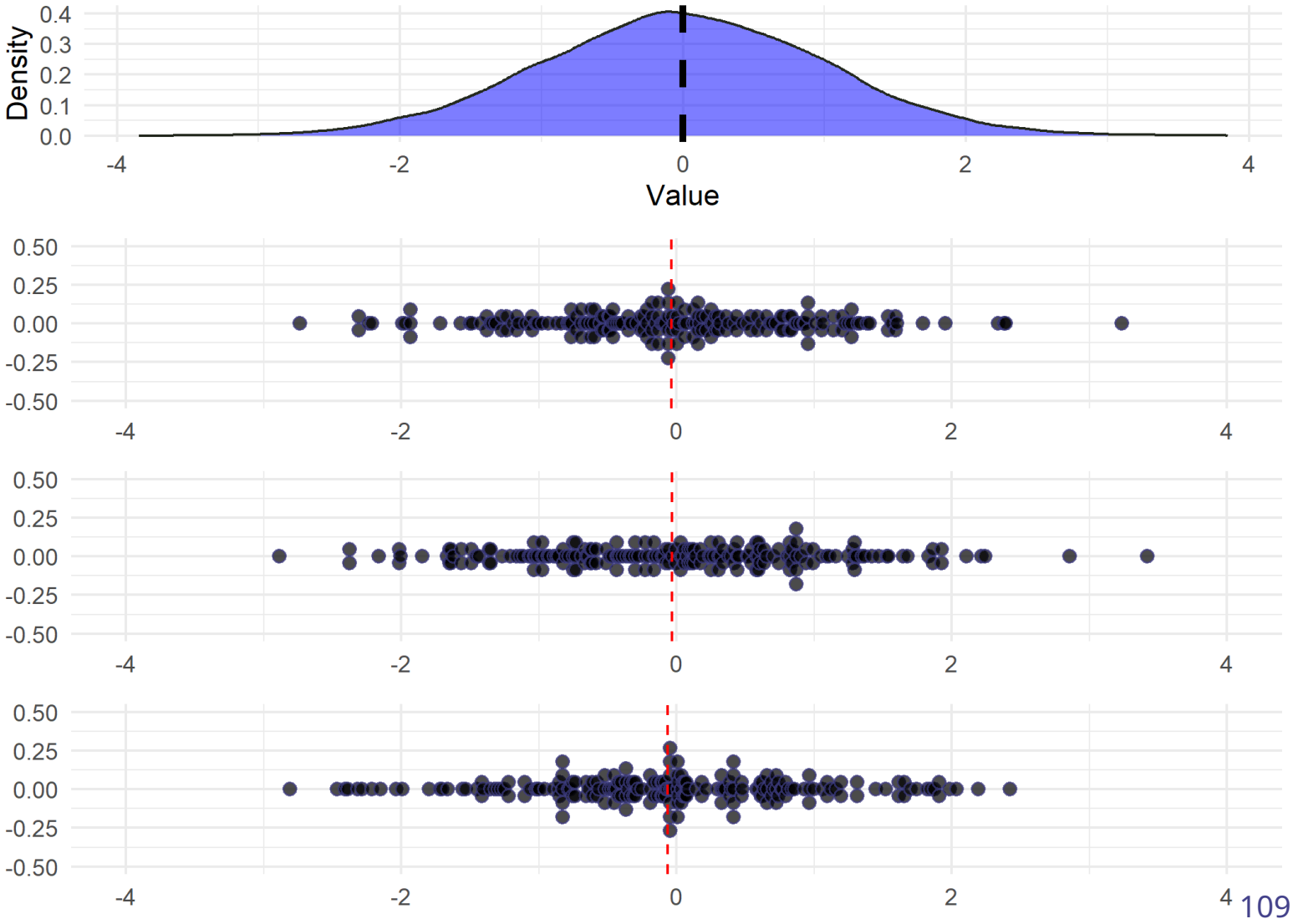
- **Intuition** - observed values usually fall closer to the sample mean than to the population mean. Distances are artificially small.



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

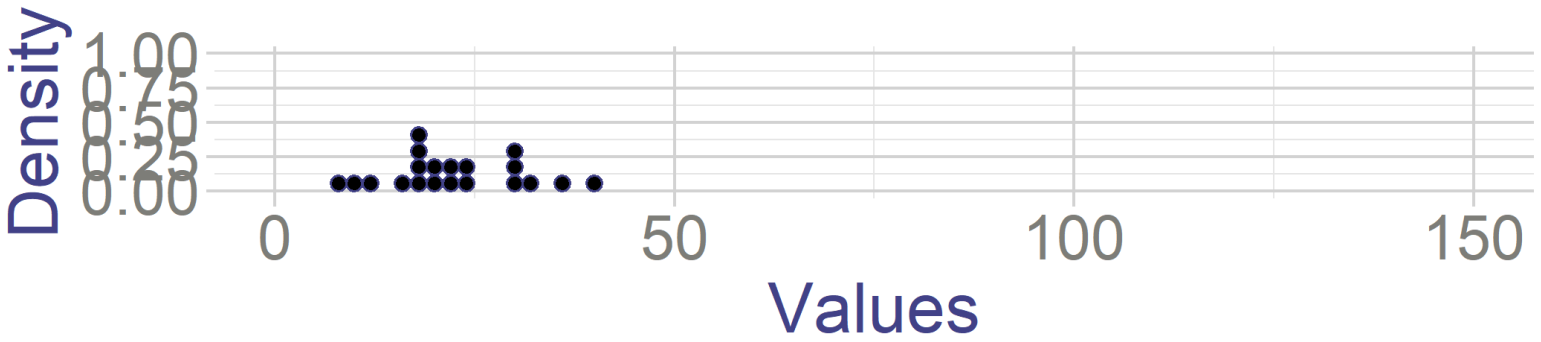
- **Intuition** - in big samples, our estimate of the population mean is already good, no need to correct



Standard Deviation

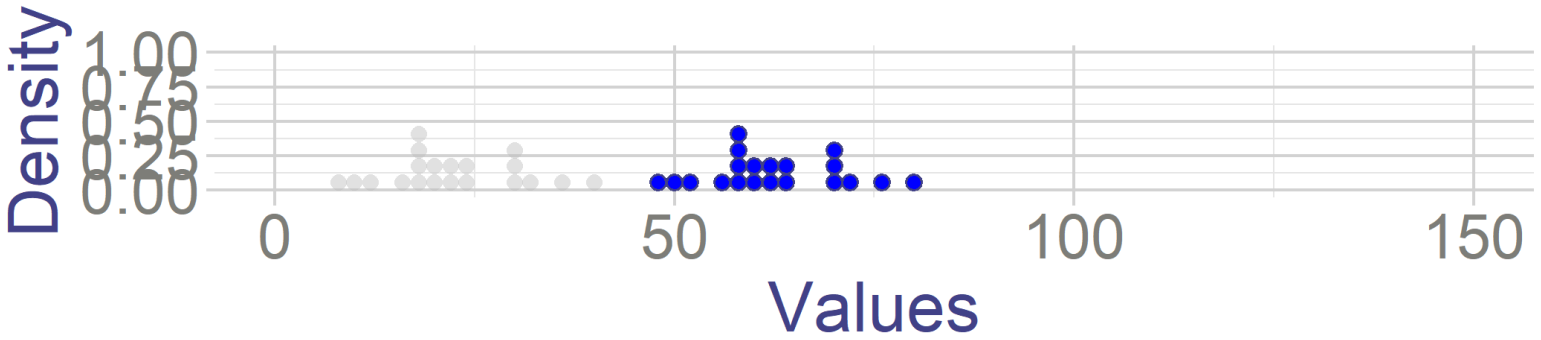
Consider a random variable X with $E(X) = \mu_x$ and standard deviation σ_x.

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?
- $Y = X + 40$



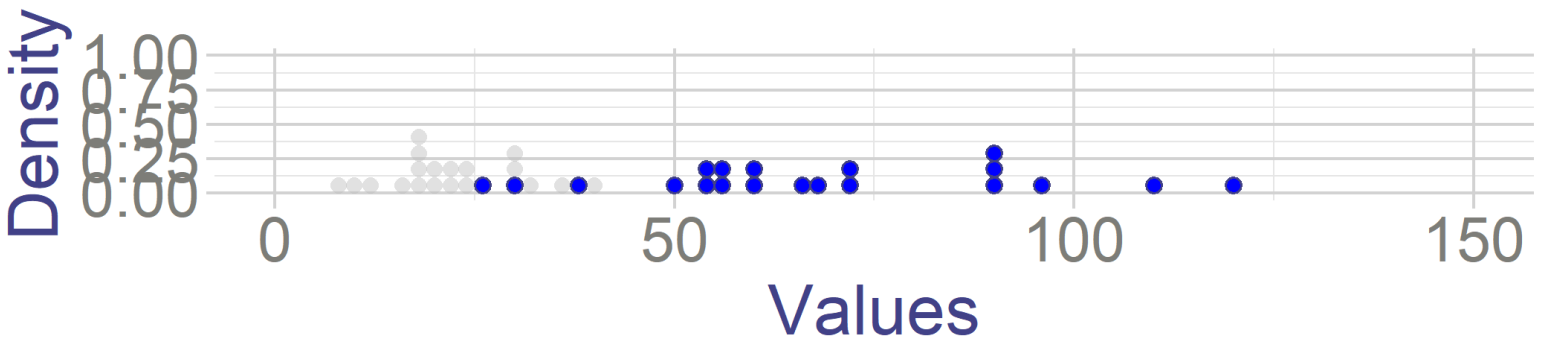
$$\mu_Y = E(Y) = \frac{\sum_{i=1}^{N}(y_i + 40)}{N} = \frac{\sum_{i=1}^{N}(x_i + 40)}{N} = \frac{\sum_{i=1}^{N} x_i + N * 40}{N} = E(X) + 40 = \mu_X + 40$$

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

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

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

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

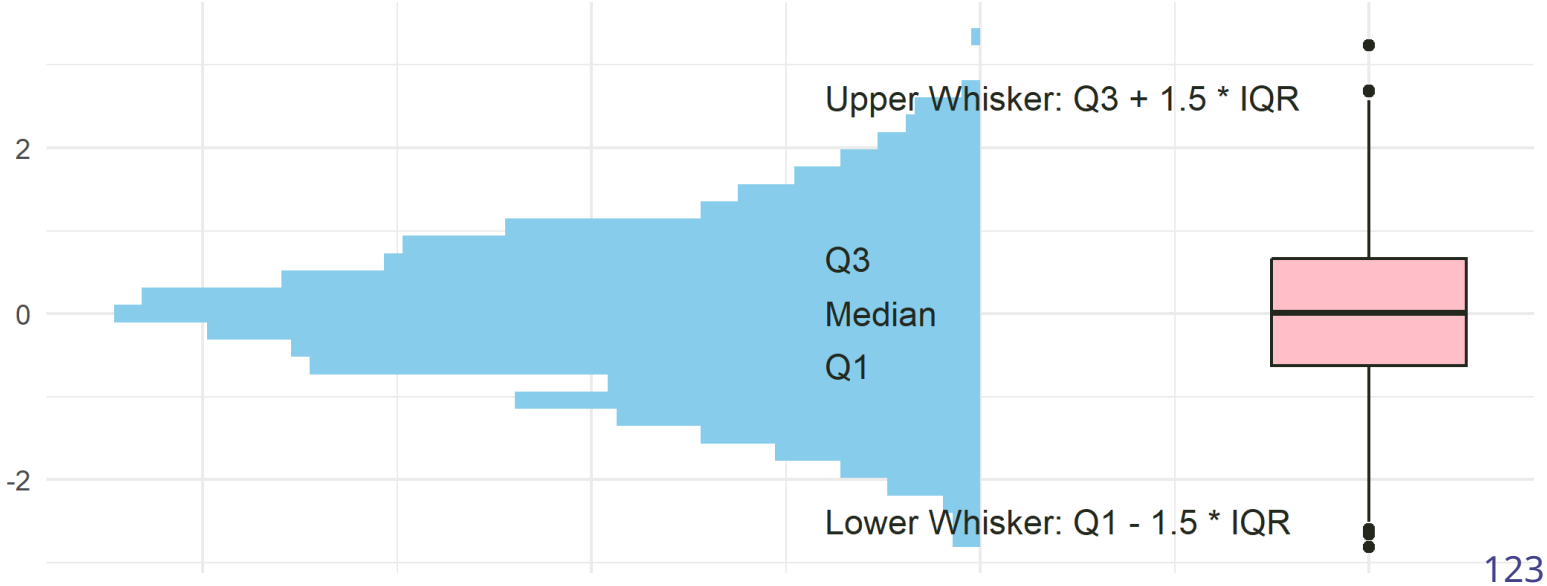
What if $y_i = bx_i + a$?

Then

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

Box and Whiskers plot

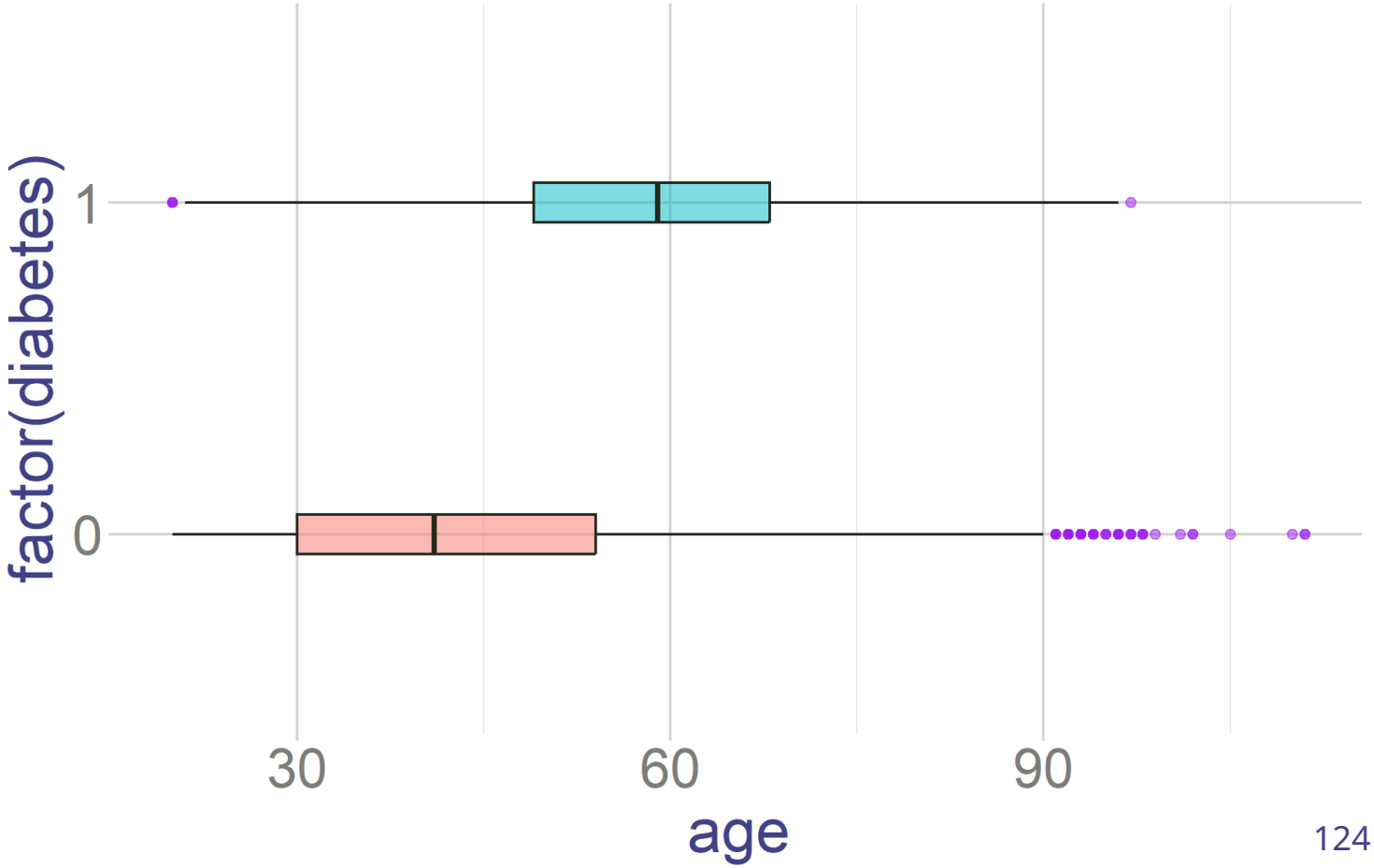
- Helps to see the distribution of the data
- Helps to see to see the outliers
  - Outliers are useful to see anomalies and potential errors in data collection
  - Whisker can be maximally 1.5 times the interquartile range
  - Any point beyond that is an outlier
  - If no point beyond 1.5 times the interquartile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



Box and Whiskers plot

Dataset comparisons

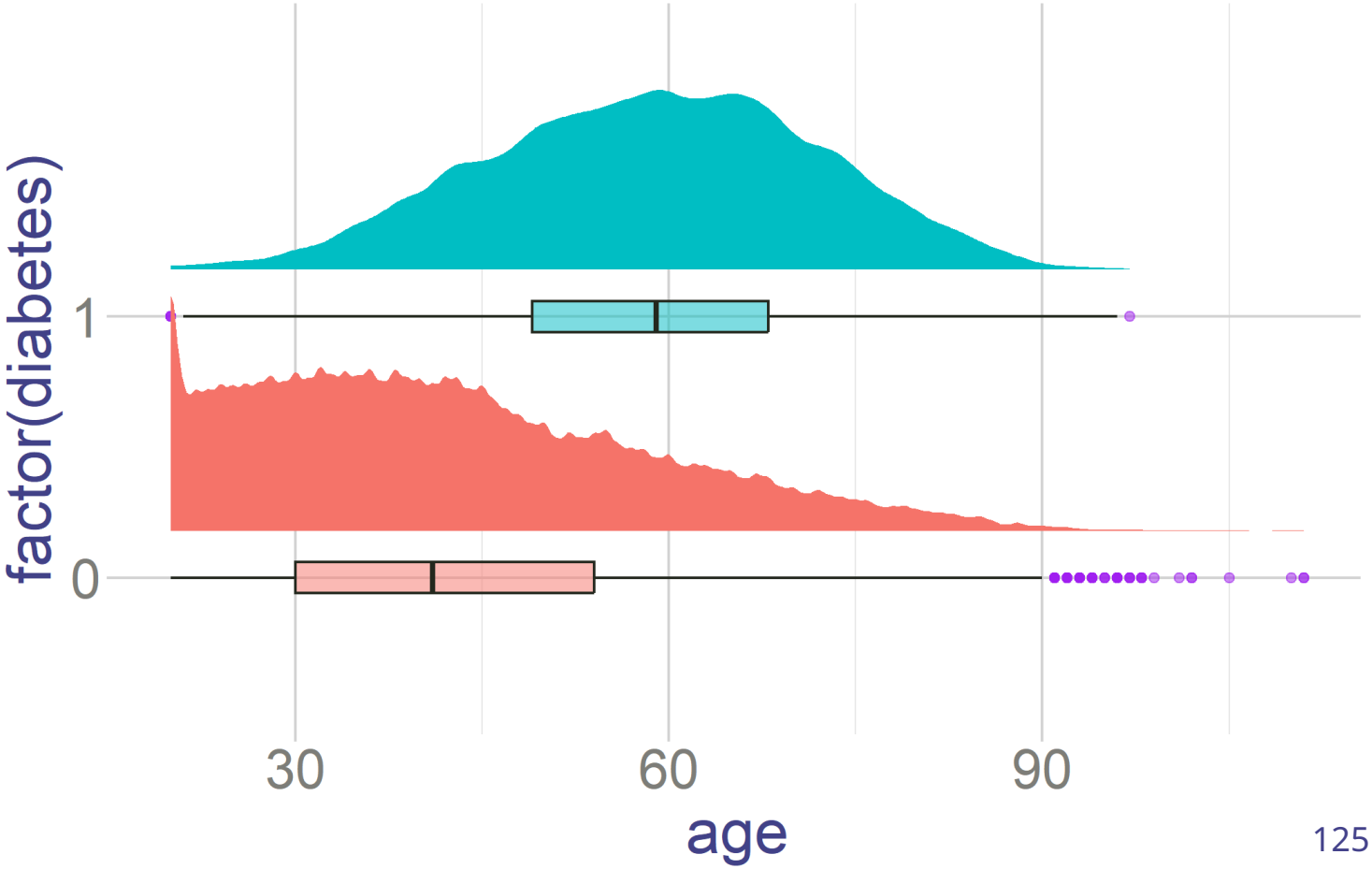
- They summarize data very well



Box and Whiskers plot

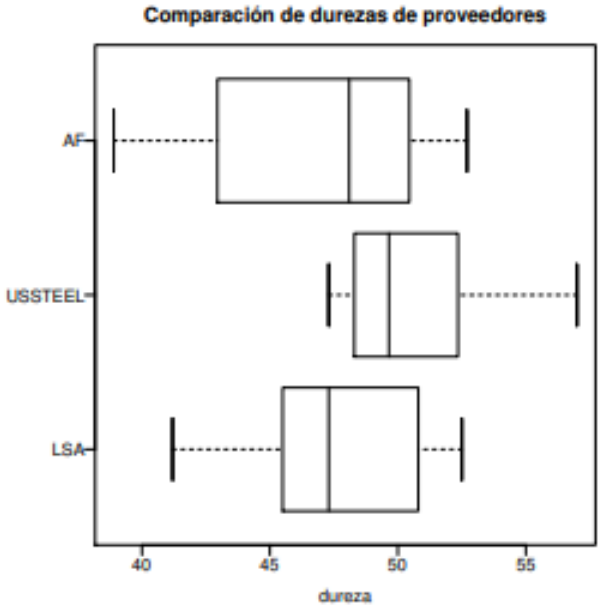
Dataset comparisons

- 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.