# Class 2b: Review of concepts in Probability and Statistics

Business Forecasting

# Summarizing Data

## Summary Statistics

# 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 $\mu$ is the sum of all observations divided by the total population size:

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

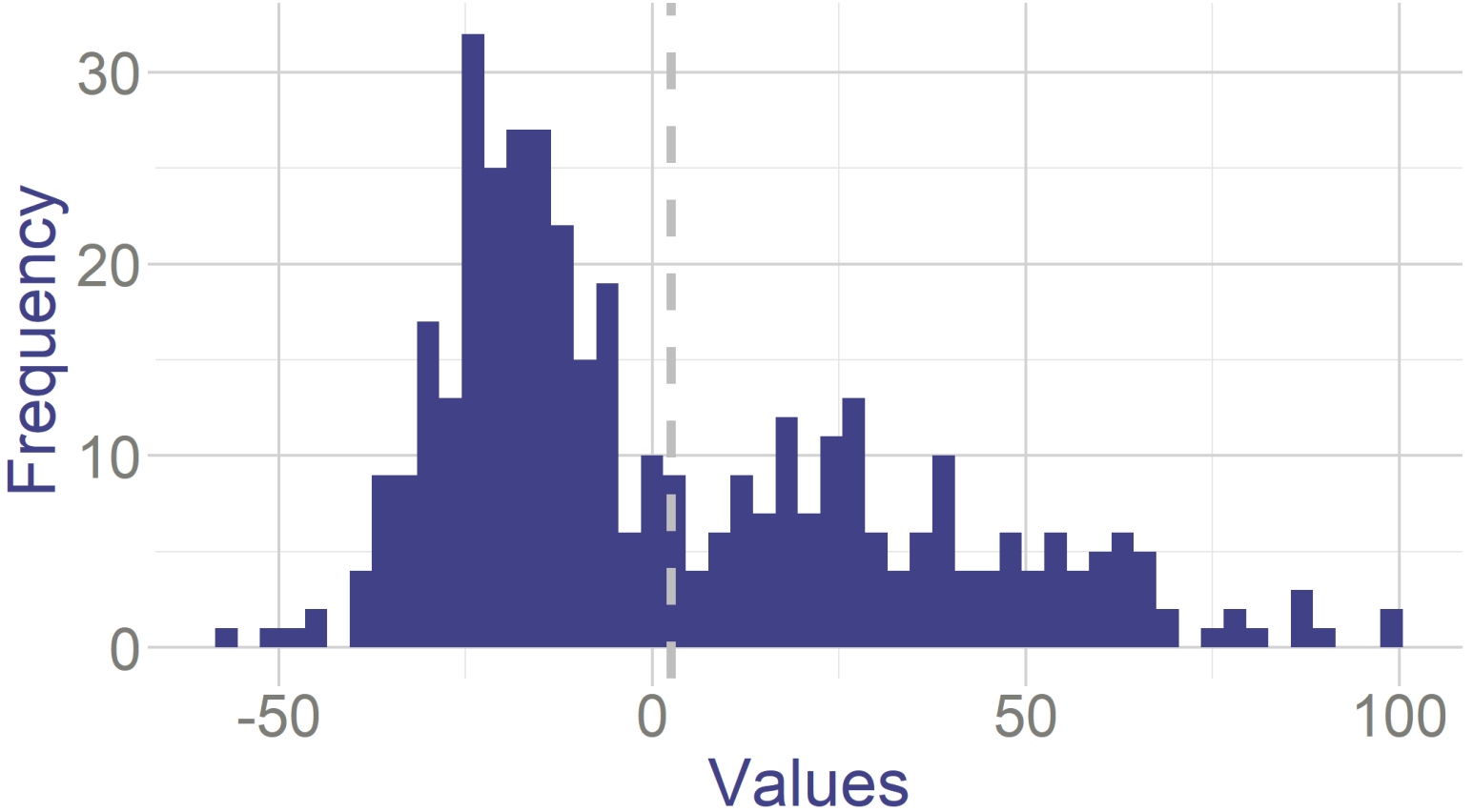
- where $N$ is the total population size, and $x_i$ are individual data points.
- The sample mean, denoted as $\bar{x}$, is the sample equivalent:

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

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:

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

# Aggregated Data

- We want to know average individual income in Mexico City
- But we only know averages by neighborhood, no individual 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

# Unweighted Mean vs. Weighted Mean

- Unweighted mean is: 25916.67 USD
- Weighted mean is: 21760.42USD
- Which one reflects average population income in CDMX?

Let

- $\mu$ be the average individual income
- $x_i$ be income of person "i"
- $N$ be the total Population in CDMX
- $N_z$ be the population in a neighborhood "z"
- $\mu_z$ be the average income in a neighborhood "z"

$$\mu = \underbrace{\frac{\sum_{i=1}^{N} x_i}{N}}_{\text{Average Individual Income}} = \frac{\sum_z \sum_{i=1}^{N_z} x_i}{\sum_z N_z} = \frac{\sum_z \frac{N_z}{N_z} \sum_{i=1}^{N_z} x_i}{\sum_z N_z} = \frac{\sum_z N_z \sum_{i=1}^{N_z} \frac{x_i}{N_z}}{\sum_z N_z} = \underbrace{\frac{\sum_z N_z \bar{x}_z}{\sum_z N_z}}_{\text{Weighted Average of Neighborhood Incomes}}$$

# 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:

$$\bar{x} = \frac{10 + 20 + 30 + 40 + 50 + 60000}{6}$$
$$= 100025$$

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.

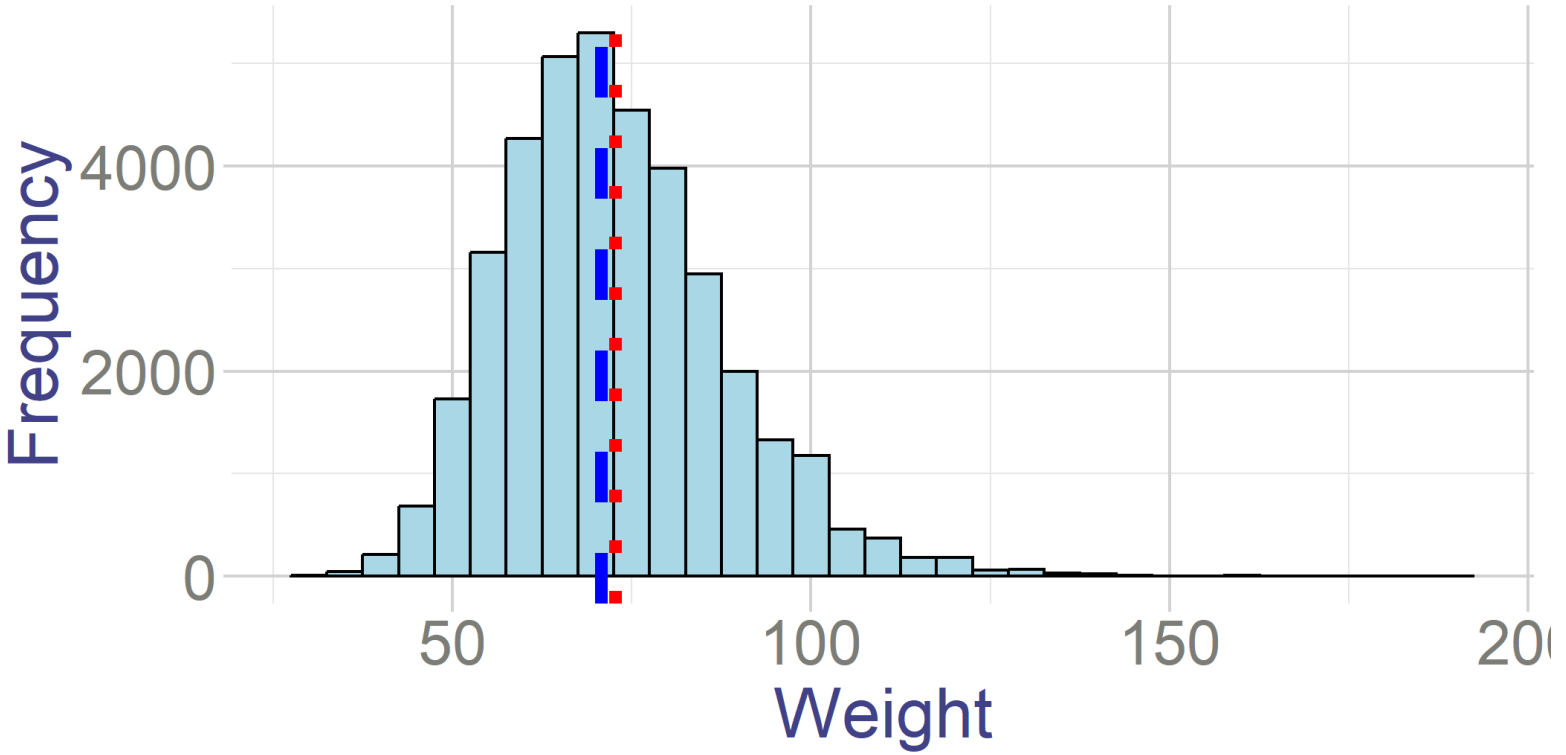
| Day | Number of Customers |
|---|---|
| 1 | 20 |
| 2 | 18 |
| 3 | 25 |
| 4 | 22 |
| 5 | 30 |
| 6 | 21 |
| 7 | 27 |

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

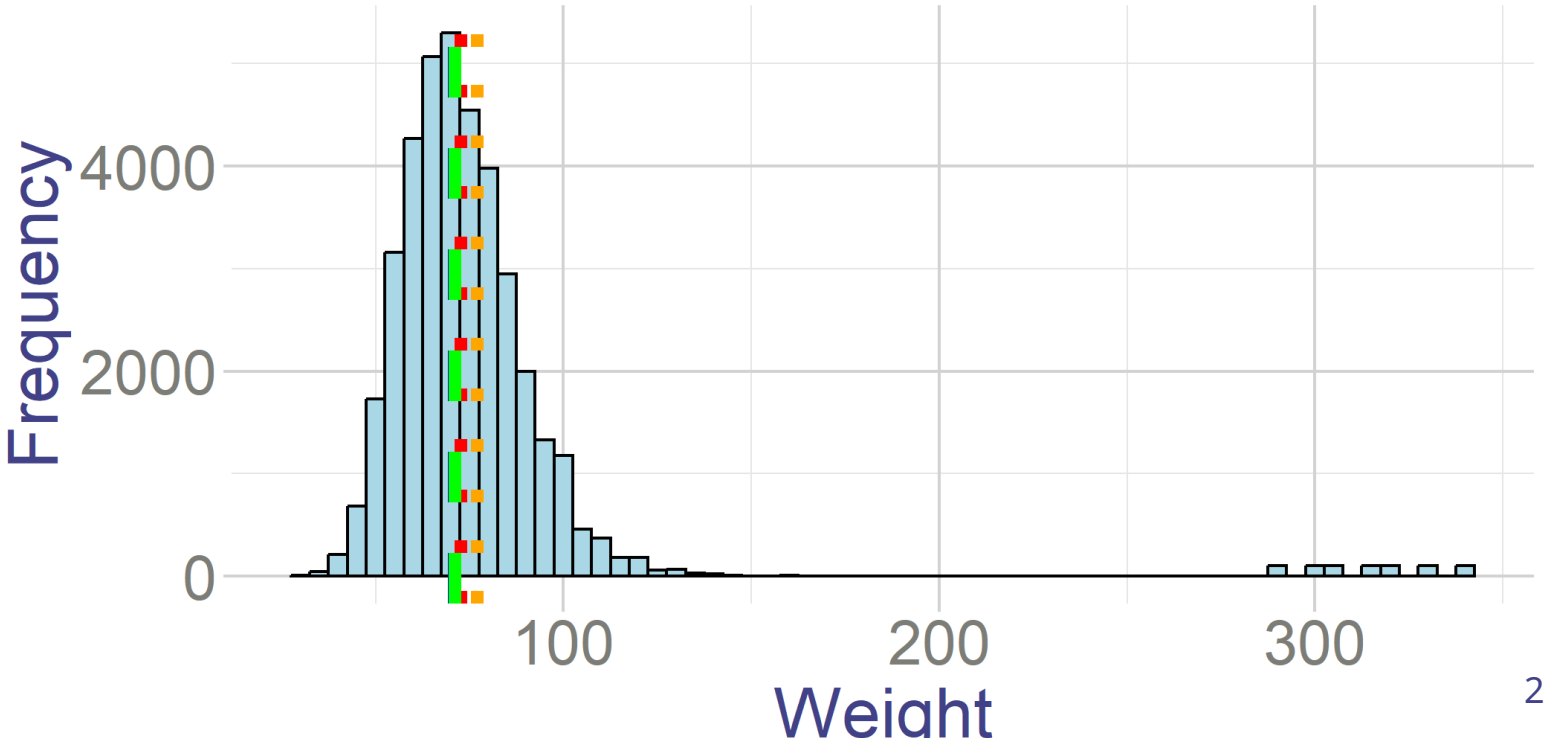
- Mean: 72.66451
- Median: 70.7536



# 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

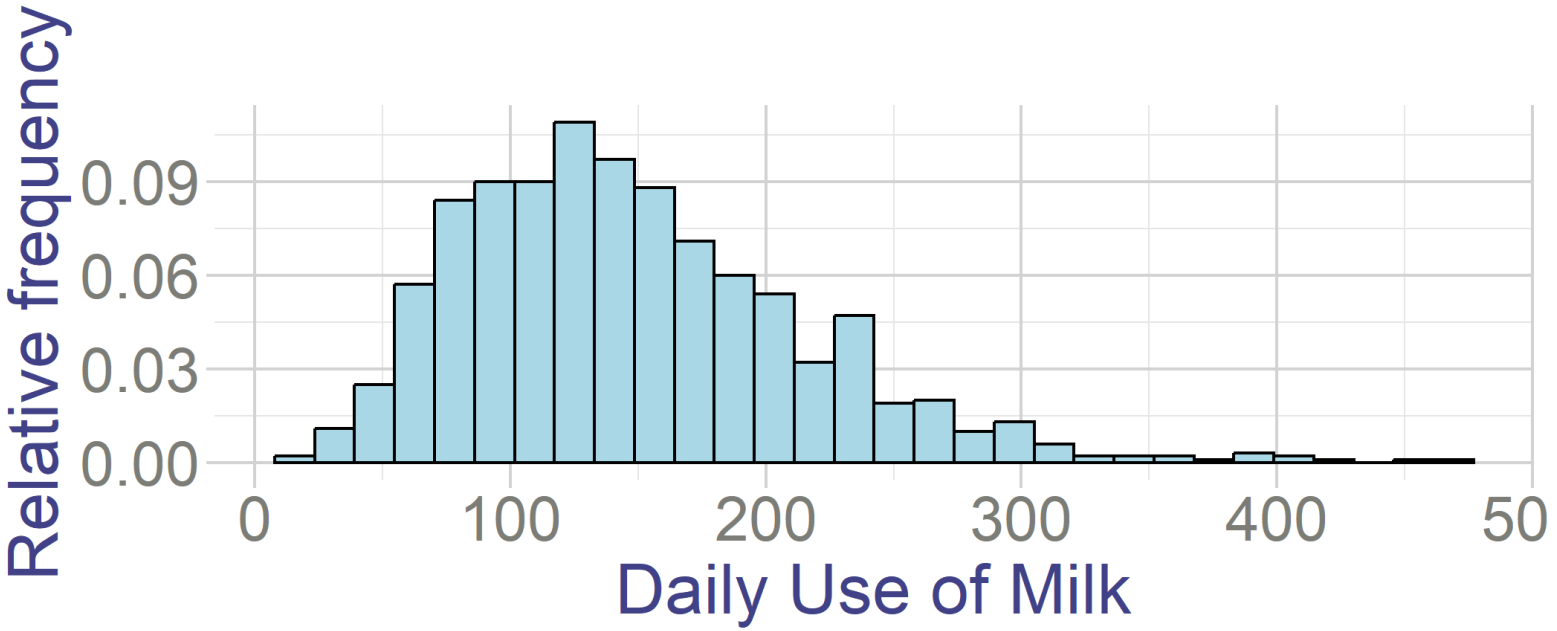
Previous 1 2 3 4 5 ... 13 Next

# 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

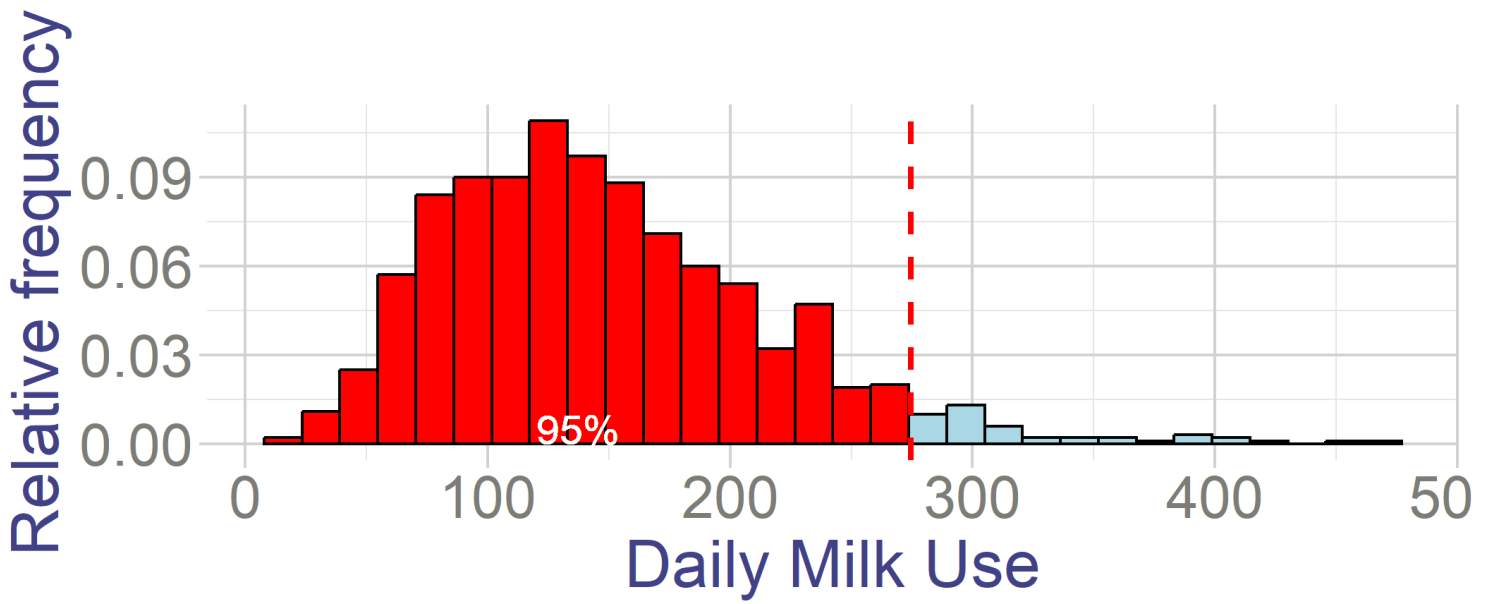
- How much inventory of milk you need to keep in your Starbucks?
- What is the tradeoff of keeping too much vs too litle inventory?
- Suppose we want to have enough of milk to cover sales on 95% of days
- To figure it out, let's look at the distribution of the daily use of milk



# 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.
- $p$th 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

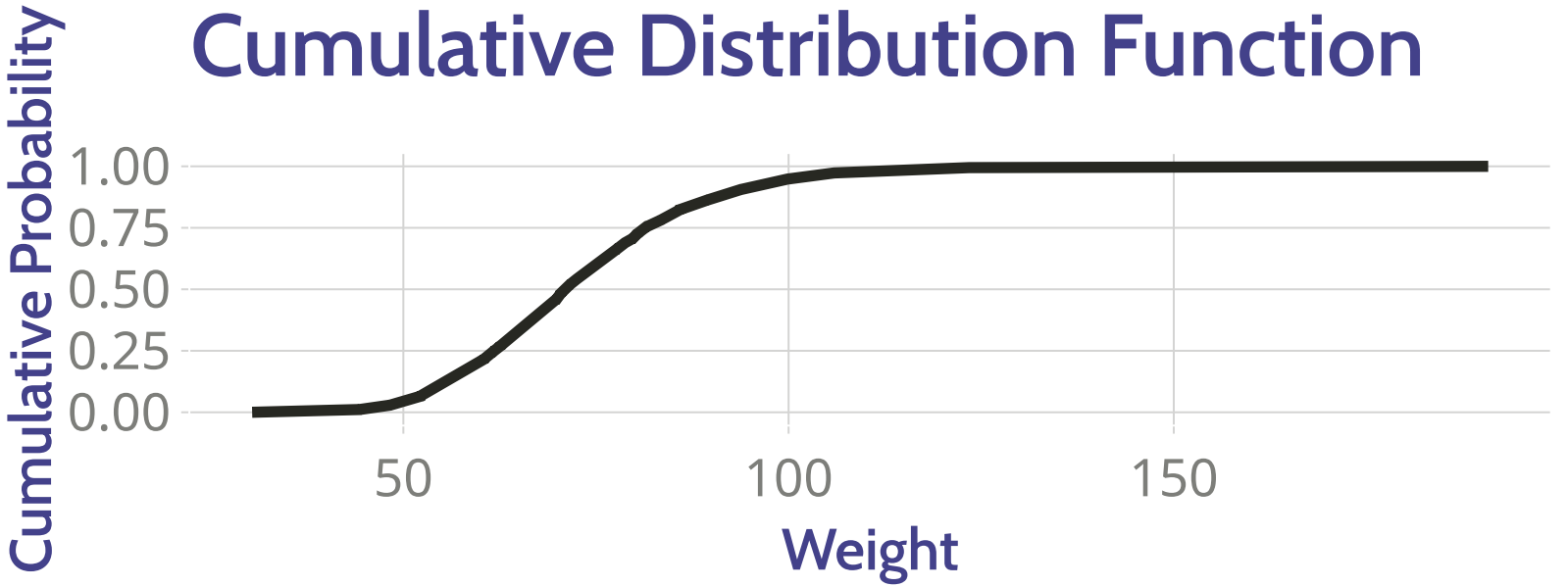
- What is the the height such that 75% of ITAM students are smaller than this height?
- What is the income level such that 25% of people in Mexico earn less than that level?
- What is the age, such that 50% of people die before that age?

# 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

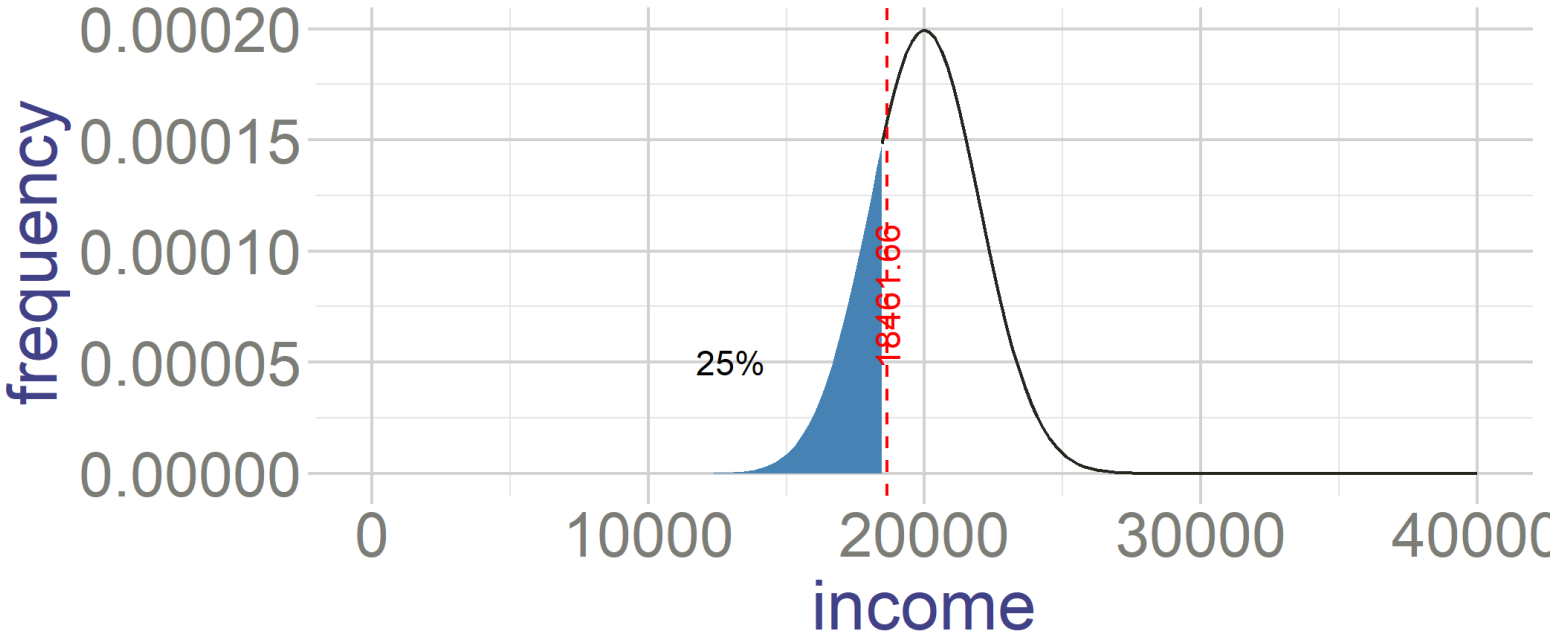
# Or use the CDF

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

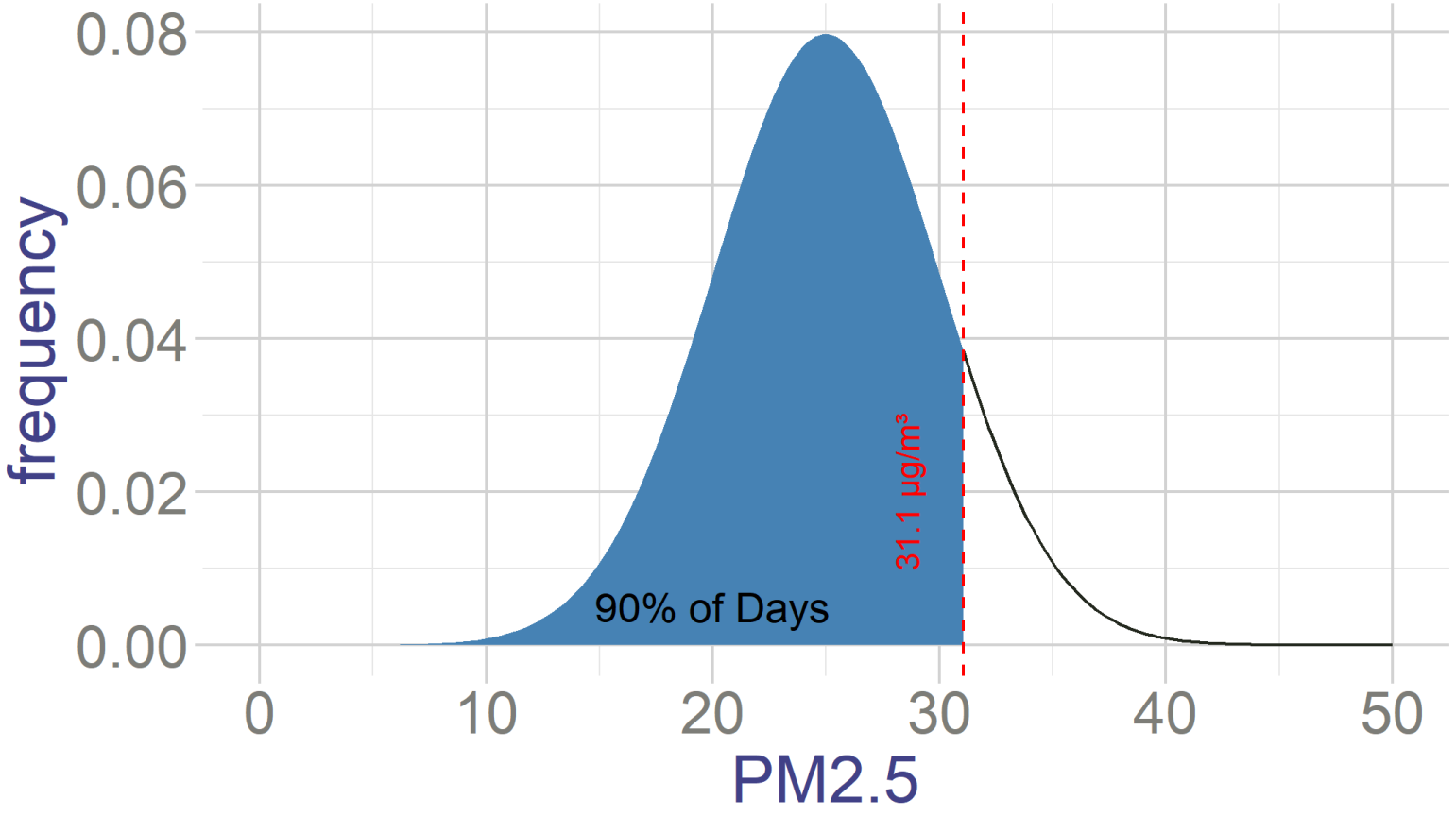


# Common values

- **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

# Exercises:

- Review Exercises:
  - PDF 2: 3,4,5,6,
- Homeworks
  - Lista 00.1: 3