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 about various **data types**
  - We will learn how to **summarize data graphically**
  - We will learn how to **summarize data with summary statistics**
  - We will learn about **comparisons and associations**
  - *Reference:* Forecasting Methods and Applications, chapter 2.1-2.4

Why do we care?

- 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

Specific Scenario

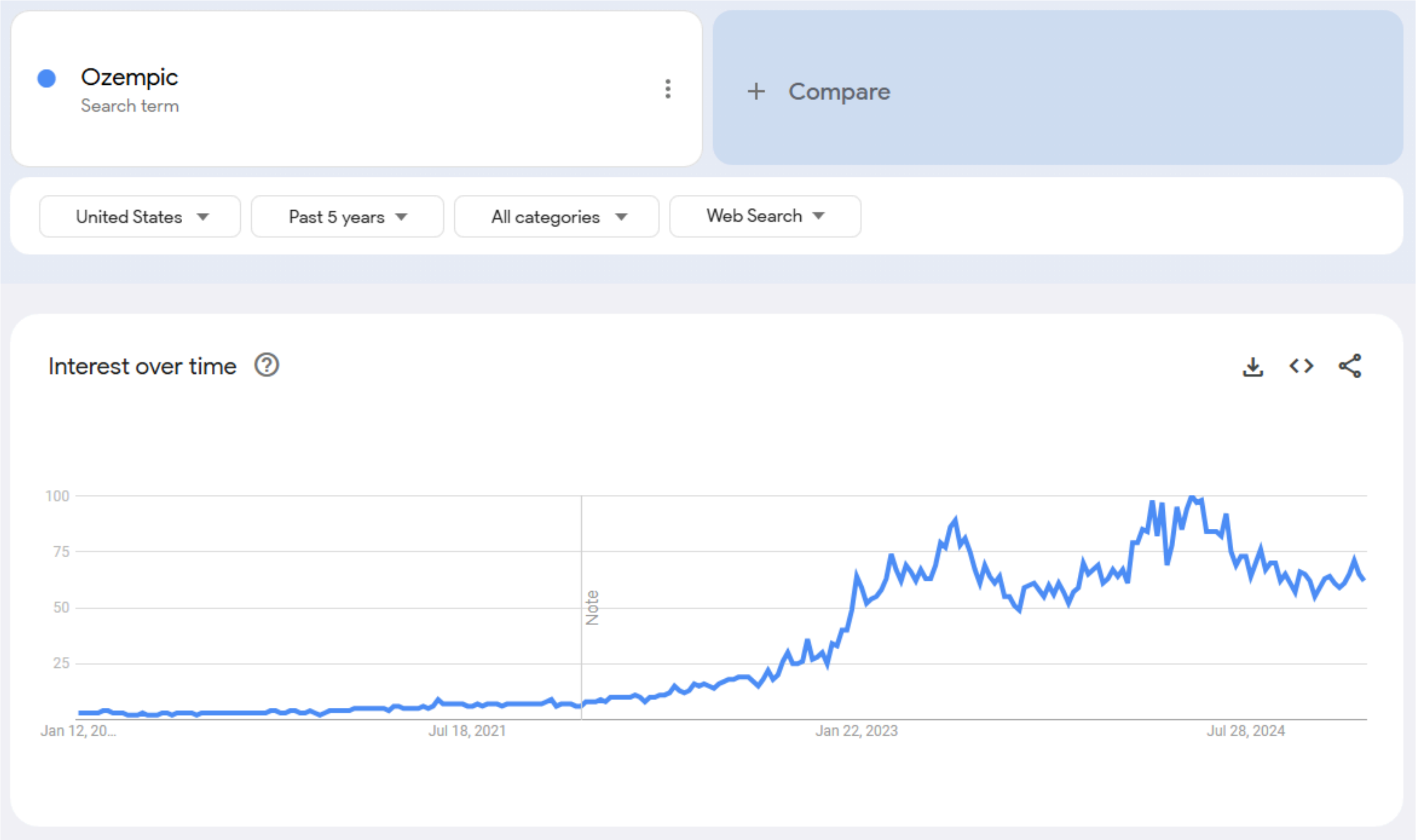
The Opportunity

- 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

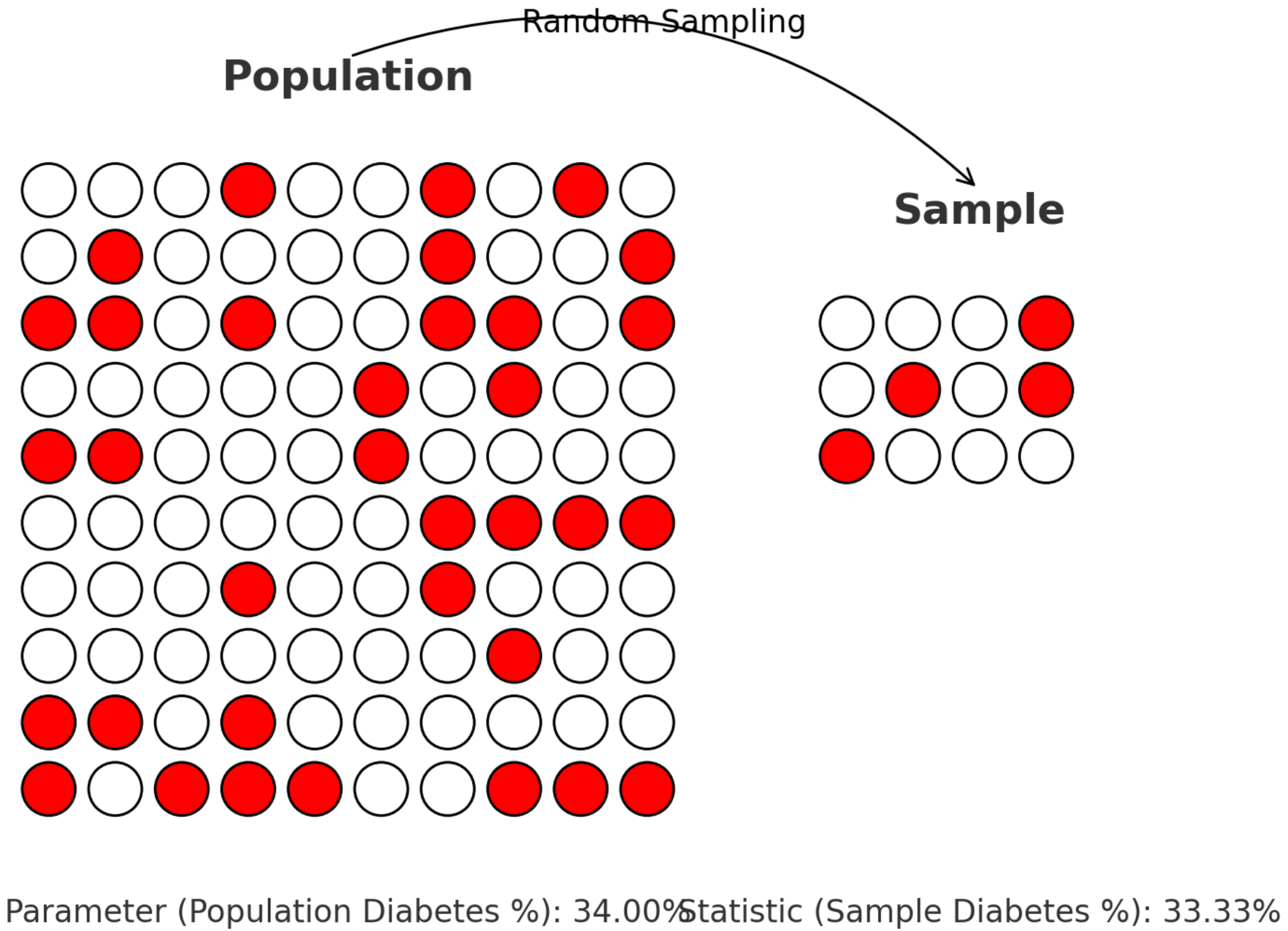
- You need to know how many people in Mexico have diabetes

  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


Random Sampling
Population
Sample
Parameter (Population Diabetes %): 34.00%
Statistic (Sample Diabetes %): 33.33%

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

- To find these numbers, you need data.
- what are types of data we can encounter?

Types of Data

Longitudinal Data

- Observations are collected for the same subject (entity) over a period of time
- Same as time series data
- Example: Tracking a company's annual revenue and number of employees over several years

Longitudinal Data Example

Show 5 entries

Year	Revenue	Employees
2018	50000	50
2019	52000	55
2020	55000	60
2021	58000	65
2022	60000	70

Showing 1 to 5 of 5 entries

Previous 1 Next

- Another Example: Share of people with Diabetes in Mexico in years 2010, 2015, 2020

Cross-Sectional Data

- Observations are collected at a single point in time
- Example: A survey of customers' satisfaction with a product and likelihood of repurchase at a certain point in time

Cross-Sectional Data Example

Show 5 entries

Customer_ID	Satisfaction_Score	Repurchase_Likelihood
1	7	Likely
2	8	Unlikely
3	5	Likely
4	9	Likely

Showing 1 to 4 of 4 entries

Previous 1 Next

- Another Example: Share of people with Diabetes in 2010 in Mexico, USA, Canada, Brazil

Panel Data

- Combines both longitudinal and cross-sectional data
- Observations are collected for multiple subjects over multiple points in time
- Example: Tracking the annual revenue and number of employees of several companies over a few years

Panel Data Example

Show 5 entries

Year	Company	Revenue	Employees
2018	A	50000	50
2018	B	52000	55
2018	C	55000	60
2019	A	58000	65
2019	B	60000	70

Showing 1 to 5 of 15 entries

Previous 1 2 3 Next

- Another Example: Share of people with Diabetes in Mexico, USA, Canada, Brazil, each country in years 2010, 2015, 2020

Q1

Show 5 entries

Month	Cryptocurrency	Market_Cap
Jan	Bitcoin	60000
Jan	Ethereum	40000
Jan	Dogecoin	10000
Feb	Bitcoin	62000
Feb	Ethereum	41000

Showing 1 to 5 of 36 entries

Previous 1 2 3 4 5 … 8 Next

Panel data

- Multiple time observation per subject (currency) and multiple subjects

Q2

Show 5 entries

Country	Population_Millions	GDP_Billions	Internet_Users_Millions
United States	331	21433	246
China	1439	15308	904
India	1380	3160	560
Brazil	213	1848	126
Russia	145	1690	116

Showing 1 to 5 of 5 entries

Previous 1 Next

Cross-sectional data

- Single (time) observation per subject (user), many subjects

Q3

Show 5 entries

Year	Electric_Car_Sales
2020	20000
2021	30000
2022	40000
2023	50000
2024	60000

Showing 1 to 5 of 5 entries

Previous 1 Next

Longitudinal data

- Multiple (time) observations of a single subject

Primary vs Secondary Data

- **Primary data** is original data collected directly from the source for a specific research purpose.
    - Experimental data (if used by team designing the experiment)
    - Survey data (if used by survey team designing the surey)
    - It is customized for a particular research objective
- **Secondary data** is data that has already been collected by someone else for a different purpose but can be used for a new research question or analysis
    - National statisics - death certificates
    - National surveys reused by researchers
    - Data from medical records
    - Data on stock market

What kind of data is it?

- Surveys: A company conducts a customer satisfaction survey to gather feedback from its customers regarding their products and services.
- Sales Reports: A business can analyze past sales data from previous years to identify trends and make strategic decisions.
- Interviews: A researcher interviews individuals to understand their opinions on a particular topic, such as political preferences or healthcare choices.
- Census Data: Government census data can be used by researchers to study demographic trends or population characteristics in a specific region
- Observations: An ecologist observes the behavior of a particular species in its natural habitat to gather data for a research project.
- Social Media Data: Companies can analyze social media posts and user engagement data from platforms like Twitter or Facebook to gain insights into customer preferences and sentiment.
- Experiments: Scientists conduct laboratory experiments to test a specific hypothesis and collect data directly from the experiments.
- Academic Journals: Researchers can review published studies and articles to gather data and insights related to their research topic.

Variable Types

Variable Types

We have two general types: Categorical and Numerical variables

Categorical Variables

- Variables that can be divided into one or more groups or categories.
  - **Ordinal:** These variables can be logically ordered or ranked.
    - *Variable:* Customer Satisfaction Survey Results
    - *Example:* Very Unsatisfied, Unsatisfied, Neutral, Satisfied, Very Satisfied
  - **Nominal:** These variables cannot be ordered or ranked.
    - *Variable:* Social Media Platforms Used
    - *Example:* Facebook, Instagram, Twitter, LinkedIn, TikTok, Snapchat

Numerical Variables

- Variables that hold numeric value and ordering is possible
    - **Discrete:** These variables can only take certain values
        - *Example*: Number of App Downloads from App Store
        - *Example*: Number of children you have
        - *Example*: Size of coke products: 0.33L, 0.5L, 1L, 2.25L

Numerical Variables

- Variables that hold numeric value and ordering is possible
  - **Discrete:** These variables can only take certain values
    - *Example*: Number of App Downloads from App Store
    - *Example*: Number of children you have
    - *Example*: Size of coke products
- **Continuous:** These variables can take any value within a range
  - *Example*: Time spent on a Webpage
  - *Example*: Exchange rate between MXN and USD
- What's the main difference between ordinal and discrete?
  - We could say 1=Very unsatisfied, 2=Unsatisfied
  - But we cannot say that very unsatisfied has half of satisfaction of person who is just unsatisfied!
  - We can order, but these numbers don't have meaning in terms of distance between them

Scales of Measurements

1. Nominal
2. Ordinal
3. Interval (Numerical)
    - No true zero point, ex: temperature
    - Differences between 5 and 10 is same as 25 and 30
    - We can't say that 10 is twice as cold as 5
4. Ratio (Numerical)
    - True zero point, ex: weight, height, age
    - If you are 20 years old, you are twice as old as your 10 years old brother

Mexican Health Survey

- Representative sample of the Mexican population n=37858.
- We will use it to investigate market for Ozempic

Show 5 entries

age	gender	weight	location_type	diabetes	Mother_diabetes	Difficulty_walking
51	Male	77.4657	Urban	0	1	A lot of difficulty
41	Female	80.0499	Urban	0	0	A lot of difficulty
44	Male	87.1874	Urban	0	1	No difficulty
68	Female	54.9827	Urban	0	0	No difficulty
52	Female	34.3283	Urban	0	0	A lot of difficulty

Showing 1 to 5 of 37,858 entries

Previous 1 2 3 4 5 … 7,572 Next

- *Age*: Numerical, Discrete
- *Gender*: Categorical, Nominal
- *Weight*: Numerical, Continuous
- *Location_type*: Categorical, Nominal
- *Diabetes*: Categorical, Nominal
- *Mother_diabetes*: Categorical, Nominal
- *Difficulty_walking*: Categorical, Ordinal

Summarizing Data

Graphical summaries

Categorical variables

Frequency Tables

Frequency table: present the absolute frequencies (counts) and relative frequencies (shares) of each category.

- Categories are mutually exclusive and collectively exhaustive
- Relative frequency of category i: $p_i = \frac{n_i}{N}$
  - n_i is count of category i
  - N is total count in the sample

Show 8 entries

Location

Category	n_i	p_i
Rural	9899	0.261
Urban	27959	0.739
Total	37858	1

Showing 1 to 3 of 3 entries

Previous 1 Next

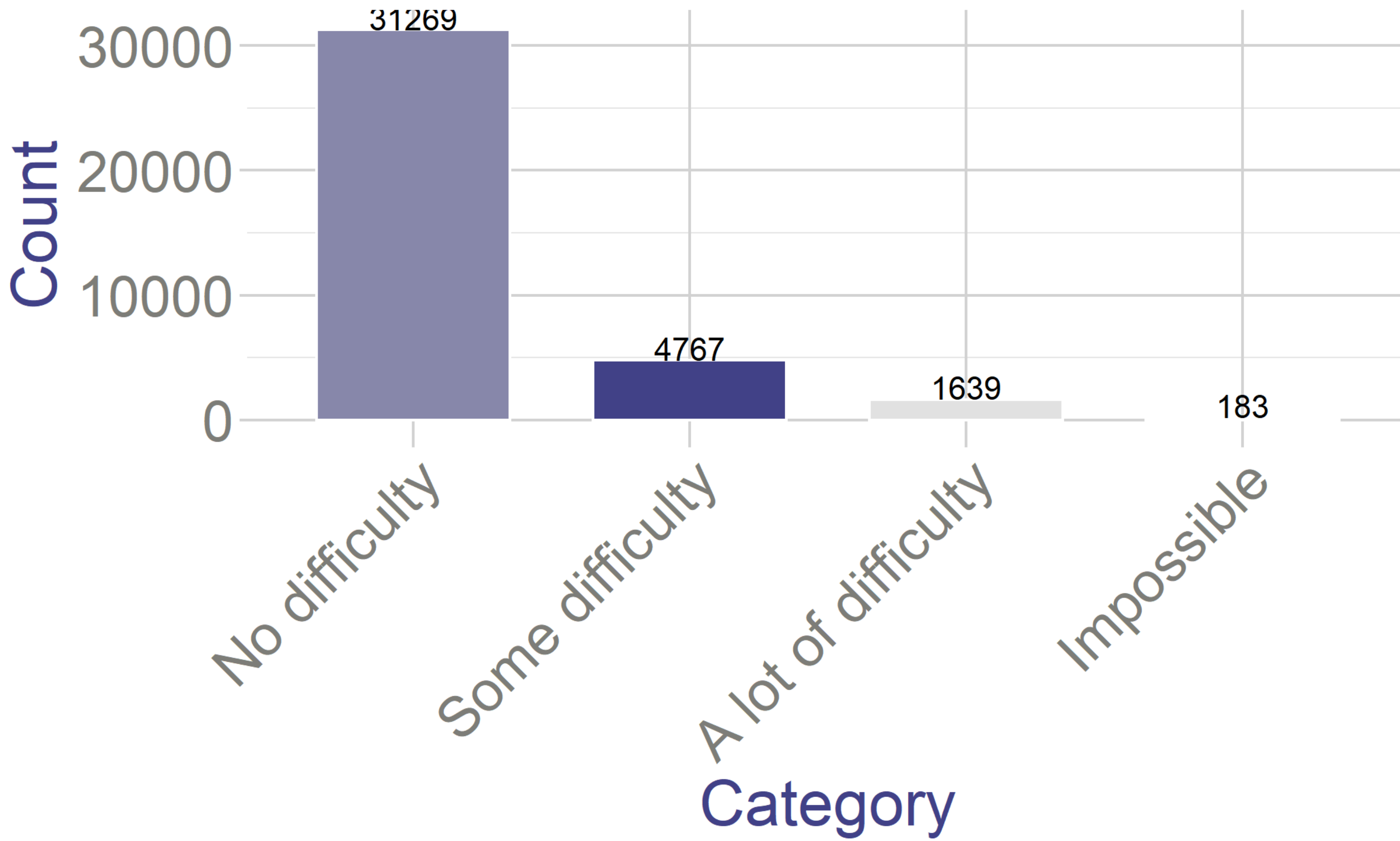
Show 8 entries

Difficulty Waking

Category	n_i	p_i
A lot of difficulty	1639	0.043
Impossible	183	0.005
No difficulty	31269	0.826
Some difficulty	4767	0.126
Total	37858	1

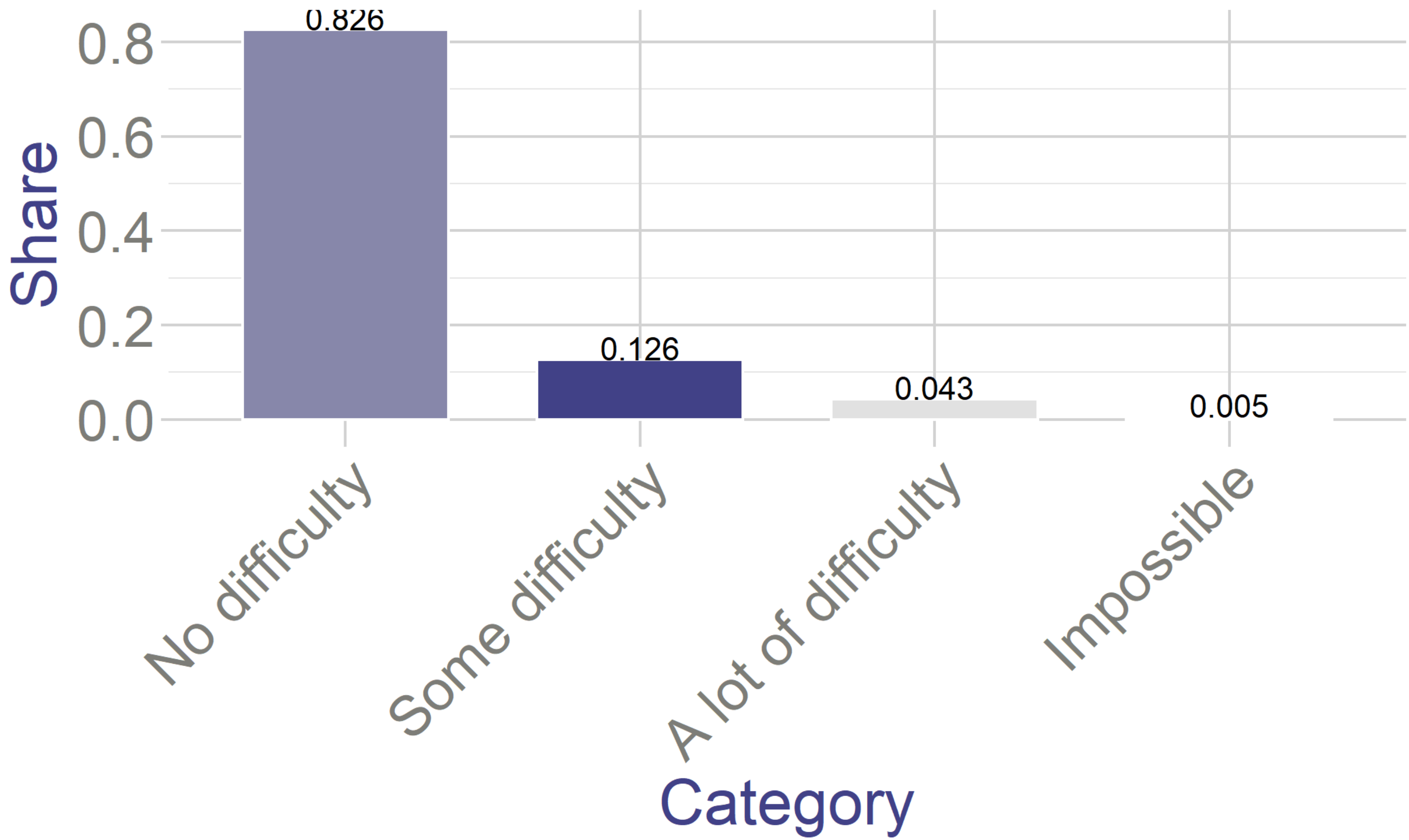
Bar Charts

Bar charts visually represents the frequency count of each category

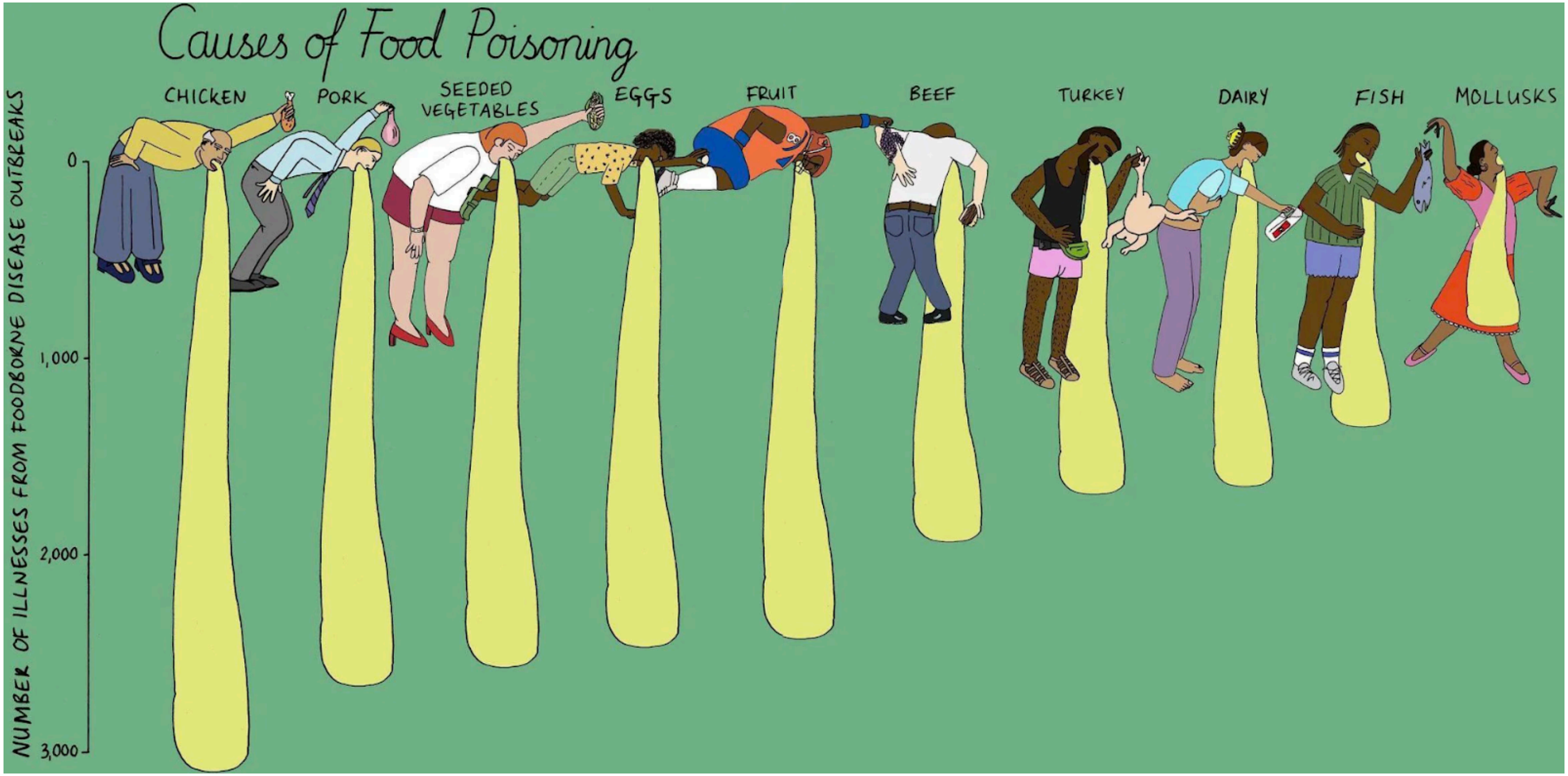


Bar Charts

Bar charts visually represents the frequency count of each category

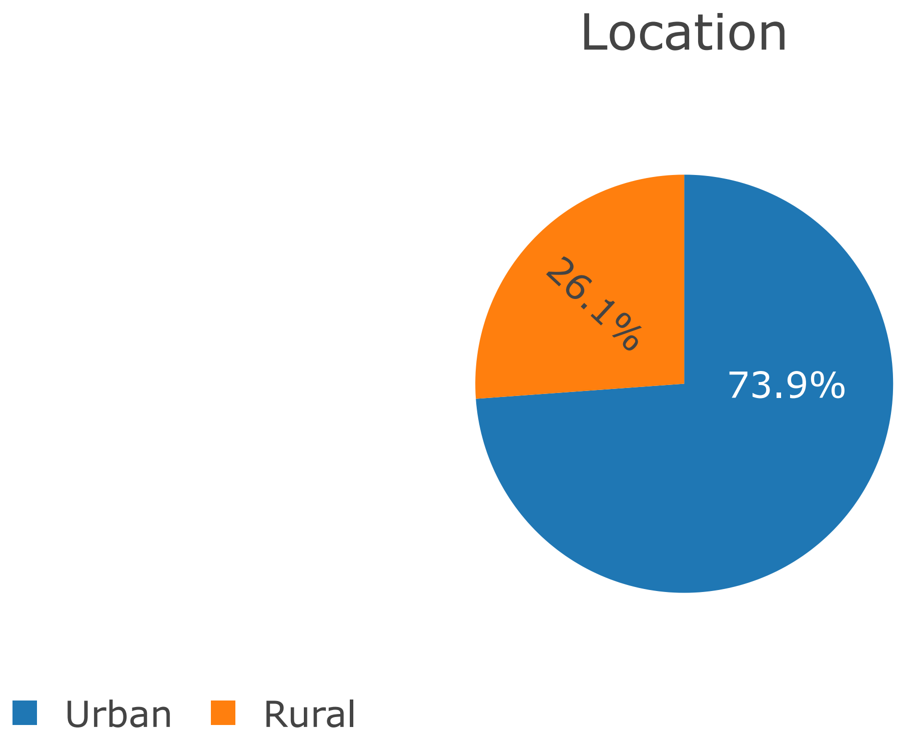


More Creative Bar Chart



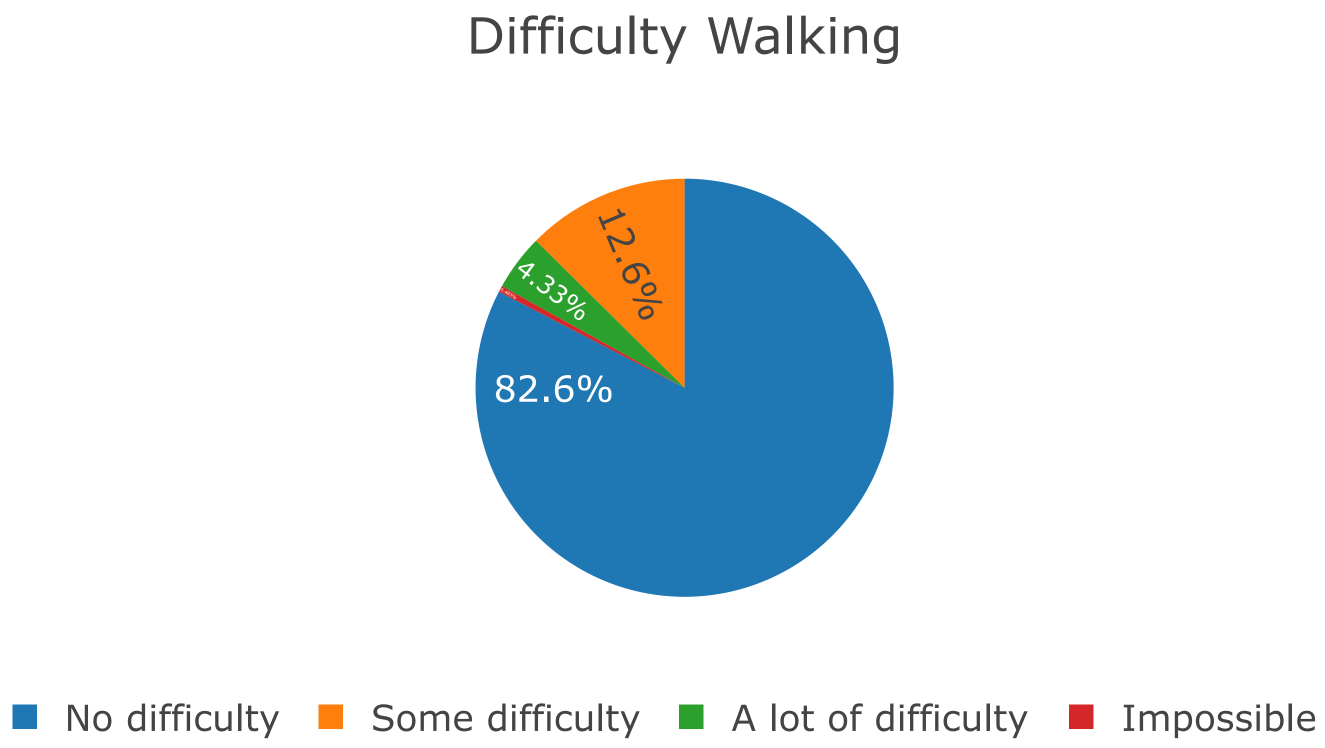
Pie Charts

Pie chart: Each slice is proportional to the category's frequency



Pie Charts

Pie chart: (Angle of) Each slice is proportional to the category's frequency



My favorite pie chart

Treemaps

Treemap: each group is represented by a rectangle, which area is proportional to its value.

Data

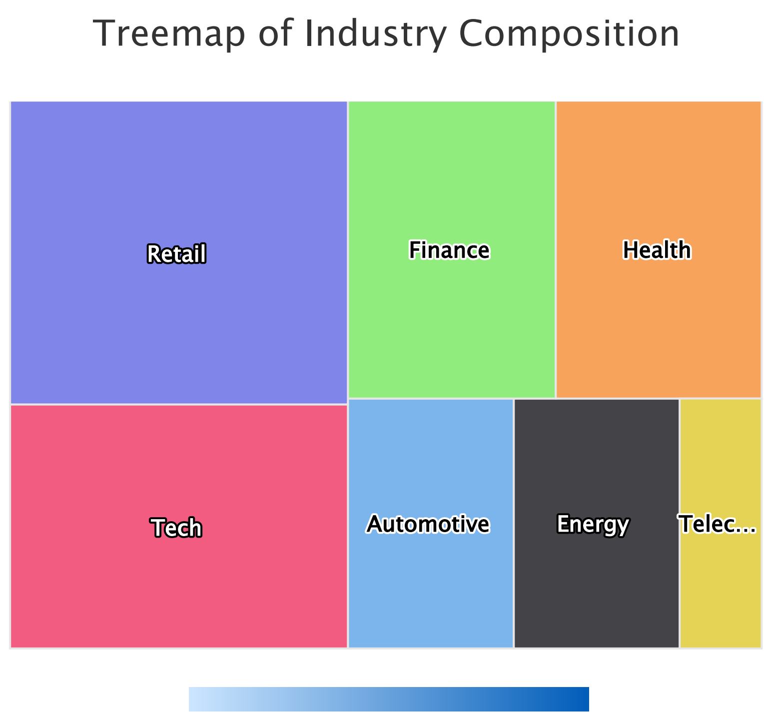
Show 8 entries

Firm	Revenue	Industry
Apple	274515	Tech
Microsoft	143015	Tech
Johnson & Johnson	82483	Health
JPMorgan Chase	142422	Finance
Alphabet	182527	Tech
Pfizer	51907	Health
Bank of America	85205	Finance
Intel	77956	Tech

Showing 1 to 8 of 20 entries

Previous 1 2 3 Next

Treemap



Numerical variables: Discrete

Dotplot: present one dot for each observation. Stacks observation of similar value

- Clearly see the distribution and the outliers
- Useless for larger data

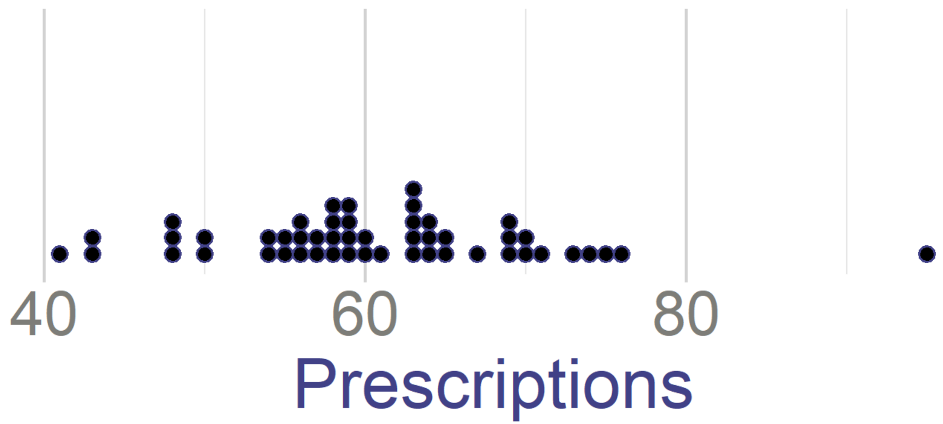
Show 2 entries

Number of prescriptions per physician

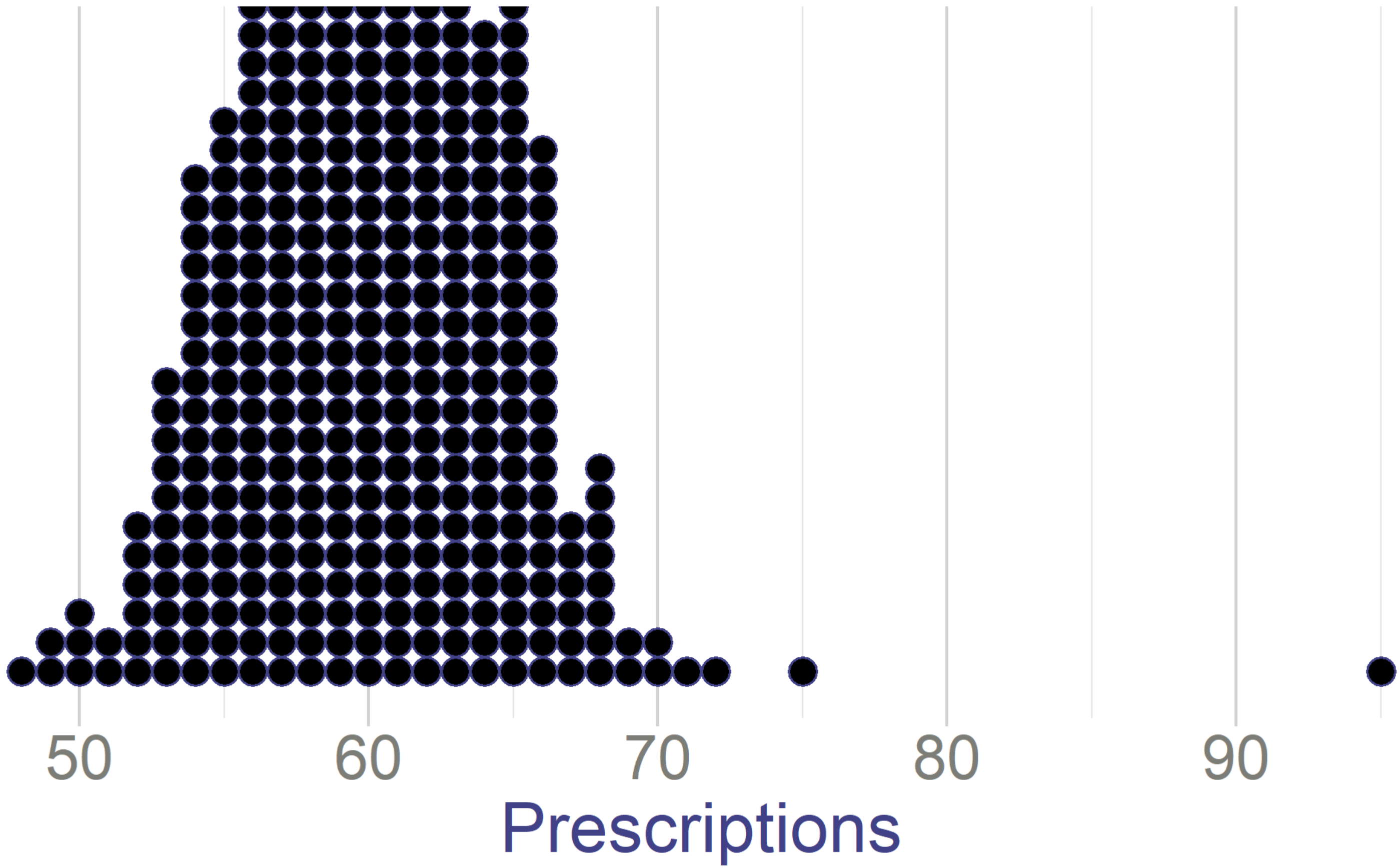
Physician	Prescriptions
Dr.1	70
Dr.2	56

Showing 1 to 2 of 50 entries

Previous 1 2 3 4 5 … 25 Next



Numerical variables: Discrete



Frequency Distribution

Suppose we survey people age 30-50 how many partners they had in their life.

- What's the distribution of partners?
- Calculate relative frequencies
- Show them on a bar graph

Data

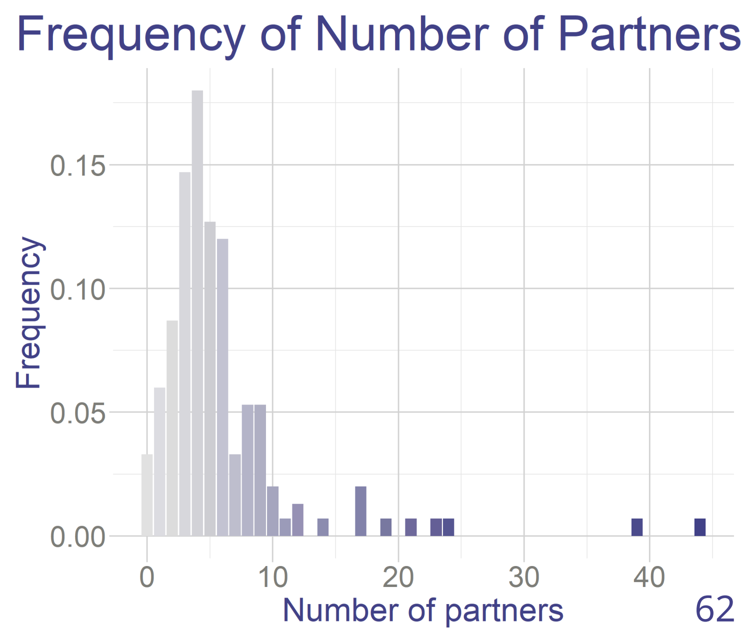
Show 6 entries

Number_of_partners	n_i	p_i
0	5	0.033
1	9	0.06
2	13	0.087
3	22	0.147
4	27	0.18
5	19	0.127

Showing 1 to 6 of 22 entries

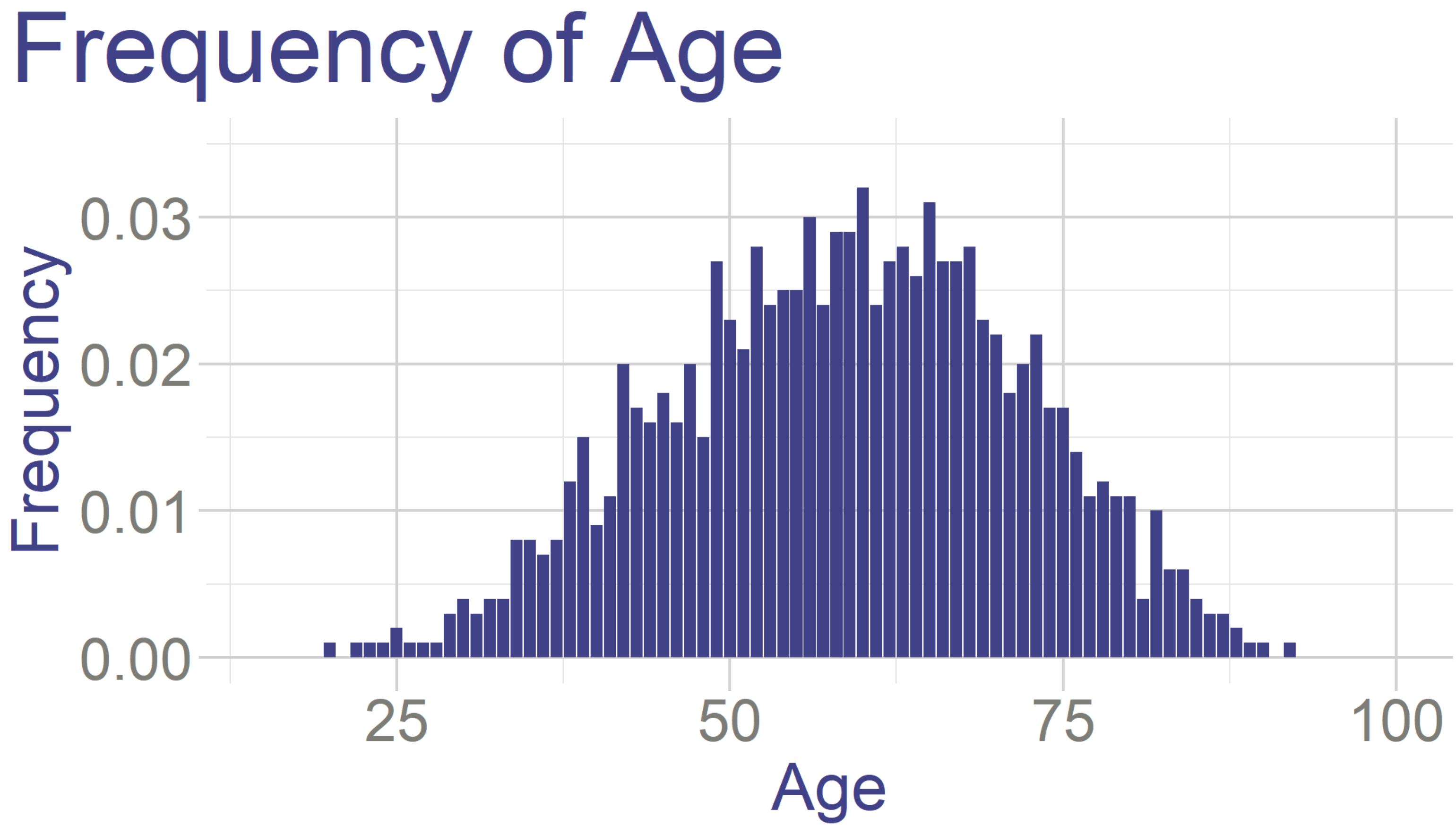
Previous 1 2 3 4 Next

Distribution



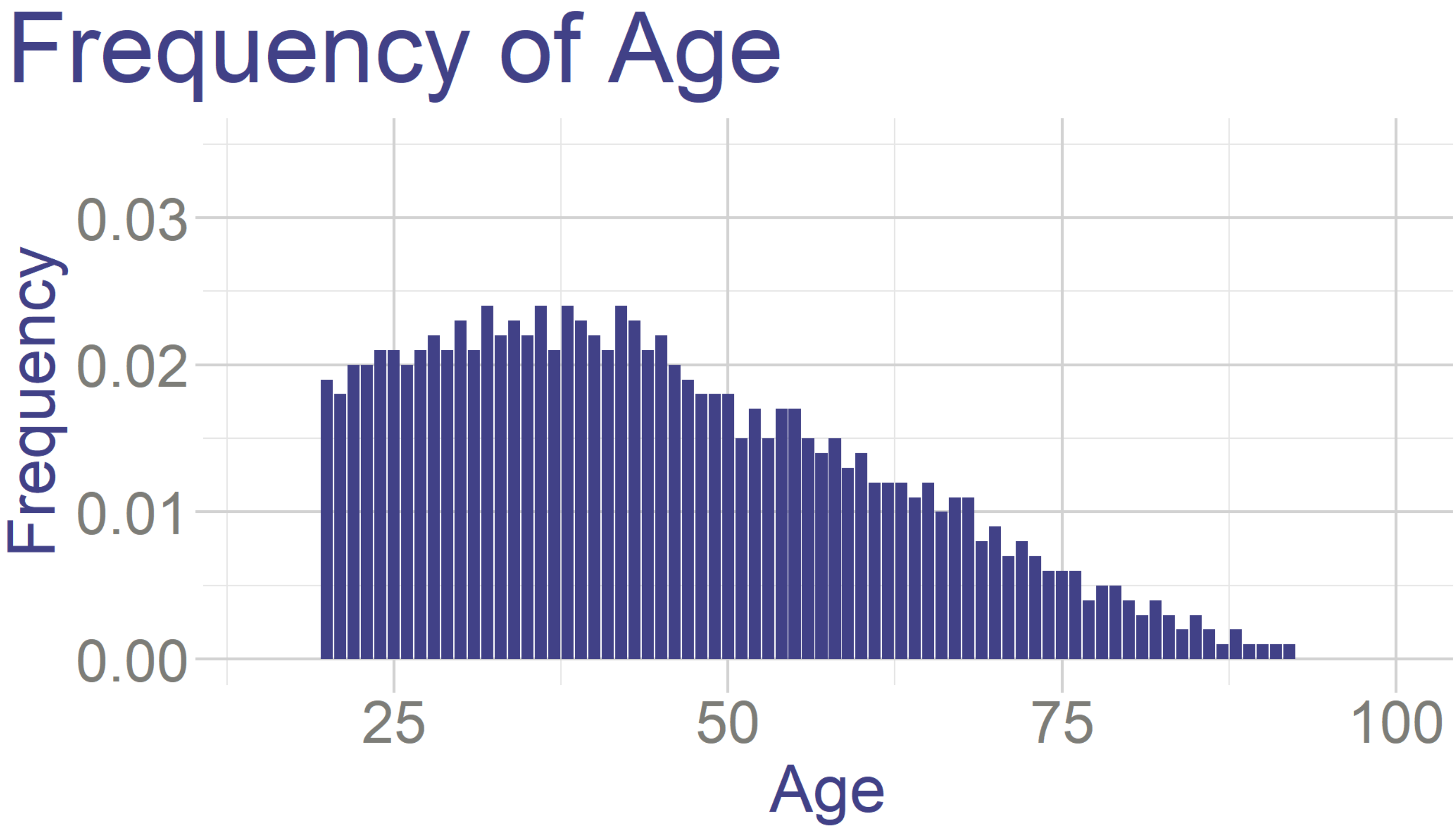
Frequency Distribution

We can also show frequency of age of people who have diabetes from our data



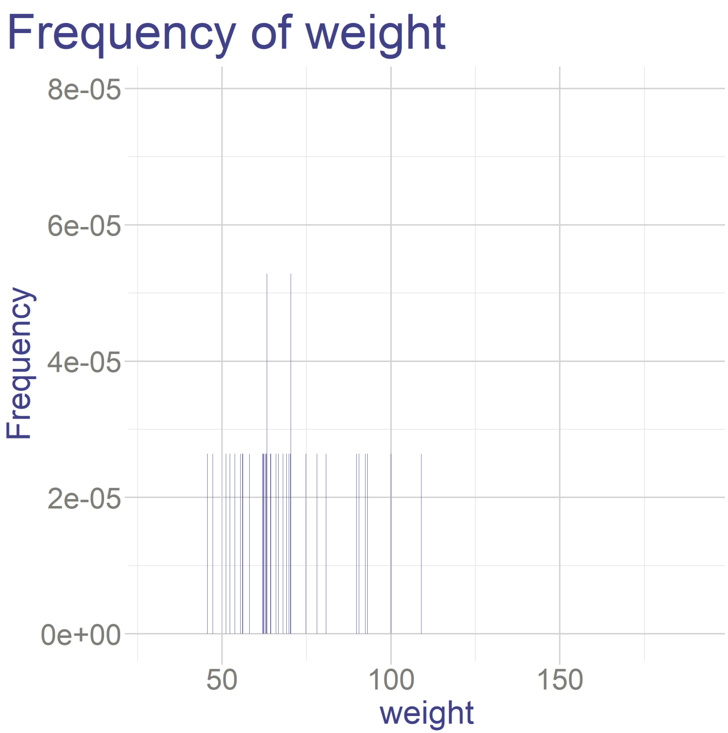
Frequency Distribution

Compare it to the age distribution in the adult population (20+)



Numerical Variables: Continuous

- What about continuous values? Why can't we do the same?



Show 6 entries

weight	n_i	p_i
30.3745	1	0.0000264
30.4593	1	0.0000264
30.5235	1	0.0000264
30.6135	1	0.0000264
30.7581	1	0.0000264
30.9106	1	0.0000264

Showing 1 to 6 of 36,297 entries

Previous 1 2 3 4 5 … 6,050 Next

- Most values never repeat, so they have very low relative frequency

Histograms

Solution: Group similar values together

- Construct intervals and show how many observations are in a given interval

Process

1. Decide how many intervals
2. And how wide they are
3. Then calculate the absolute and relative frequencies of each interval
4. Plot it with bars

My approach

- I want k (example k=5) equal intervals
- Divide the range of the data into k equal intervals
    - *Range* is max-min of the data

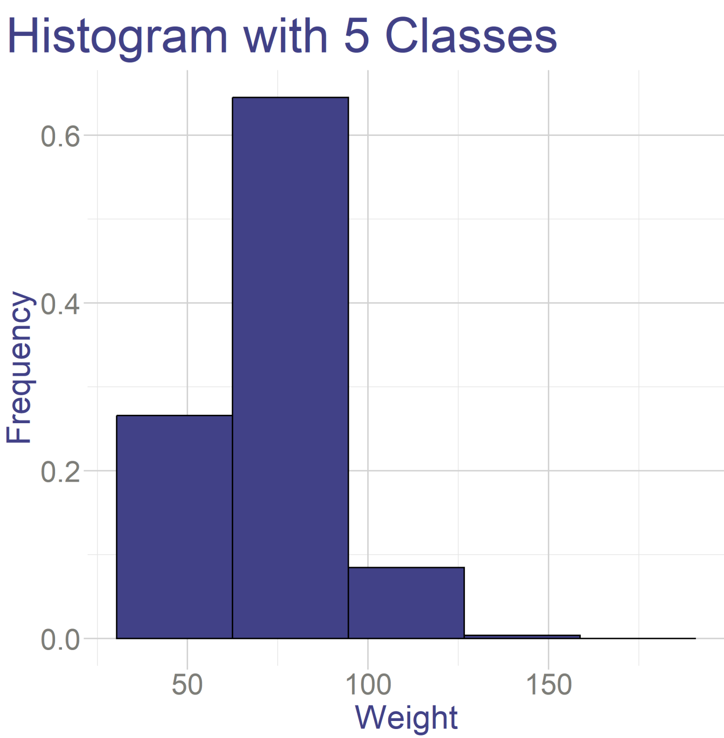
```
# Calculate max and min
max_value <- max(Health_data$weight)
min_value <- min(Health_data$weight)

# Calculate the difference
range <- max_value - min_value
```

```
## [1] "Range= 190.8078 - 30.3745 = 160.4333"
```

- With 5 intervals, each will be 32kg wide
- The first one starts at the minimum value (30.3745)
- The last one ends at the maximum value (190.8078)
- Calculate how many observations I have in each interval and what's the relative frequency

Histograms

- Midpoint represents middle of the interval - center of the bar
- P_i is cumulative frequency: share of observations in this or smaller interval
  - *Example*: $P_{(62.46-94.55)} = 0.911$
  - *Interpretation*: 91.1% of people have weight lower than 94.55kg



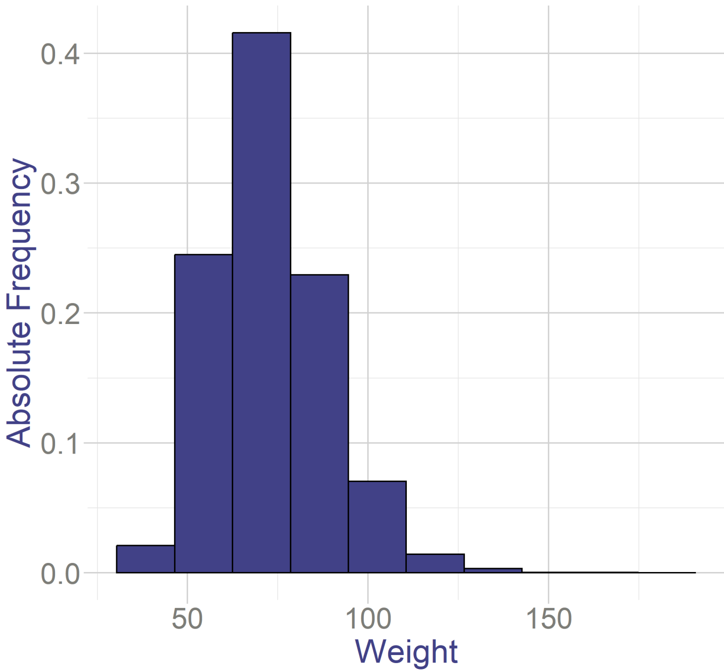
Show 6 entries

Interval	Midpoint	n_i	p_i	P_i
30.37 - 62.46	46.42	10068	0.2659411	0.2659411
62.46 - 94.55	78.5	24430	0.6453061	0.9112472
94.55 - 126.63	110.59	3206	0.0846849	0.9959321
126.63 - 158.72	142.68	143	0.0037773	0.9997094
158.72 - 190.81	174.76	11	0.0002906	1

Showing 1 to 5 of 5 entries　Previous 1 Next

Histogram with 10 Classes

Now, let's increase the number of classes to 10.

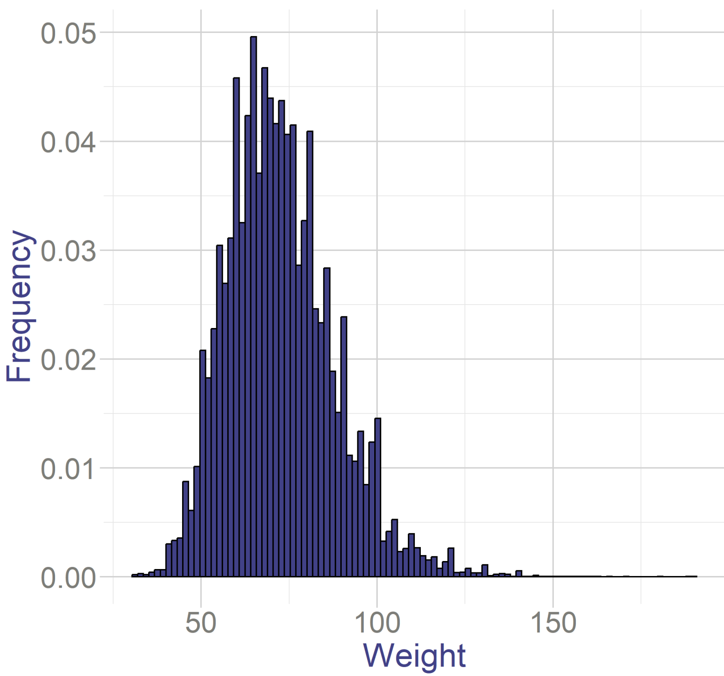


Show 6 entries

Interval	Midpoint	n_i	p_i	P_i
30.37 - 46.42	38.4	796	0.0210259	0.0210259
46.42 - 62.46	54.44	9272	0.2449152	0.2659411
62.46 - 78.5	70.48	15742	0.415817	0.6817581
78.5 - 94.55	86.53	8688	0.2294891	0.9112472
94.55 - 110.59	102.57	2661	0.070289	0.9815362
110.59 - 126.63	118.61	545	0.0143959	0.9959321

Showing 1 to 6 of 10 entries Previous 1 2 Next

Histogram with 100 Classes



Show 6 entries

Interval	Midpoint	n_i	p_i	P_i
30.37 - 31.98	31.18	8	0.0002113	0.0002113
31.98 - 33.58	32.78	11	0.0002906	0.0005019
33.58 - 35.19	34.38	7	0.0001849	0.0006868
35.19 - 36.79	35.99	16	0.0004226	0.0011094
36.79 - 38.4	37.59	24	0.0006339	0.0017433
38.4 - 40	39.2	24	0.0006339	0.0023772

Showing 1 to 6 of 100 entries

Previous 1 2 3 4 5 … 17 Next

- Helps to see the distribution and outliers
- Is more always better?
- With smaller intervals, histogram tends to the **probability density function**

Probability Density Function (PDF)

Definition

- **Probability Density Function (pdf)** describes the probability distribution of a continuous random variable.
- It **does not** give probability at a given value (this is always 0 for continous variable)
- It shows which in which intervals that variable the most often appears
- It is used to calculate the probability of the random variable being in a given interval
- Area under it always adds up to 1

Example

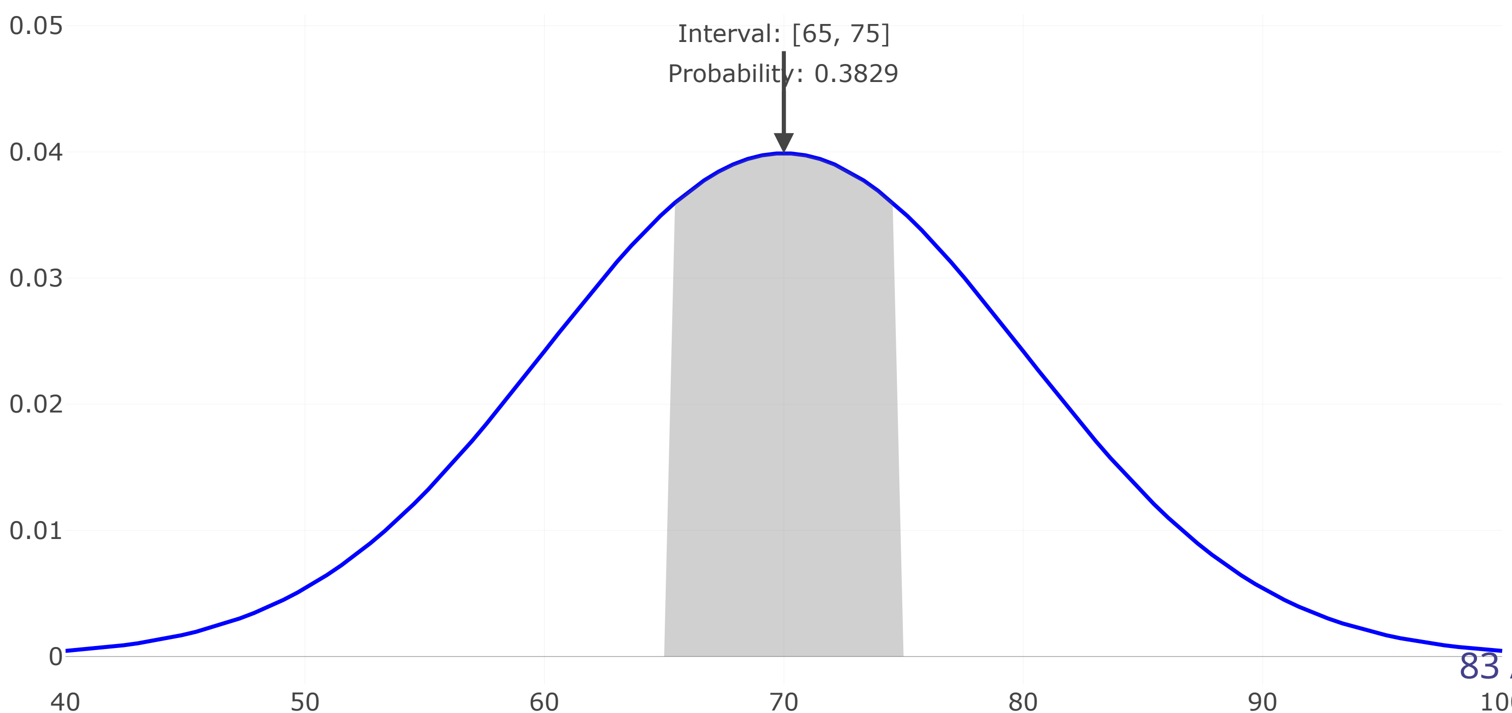
We have a random variable X representing the weight of adults in Mexican population. The PDF of X helps to describe the likelihood of finding a person of a specific weight within a range (e.g., between 58kg and 60kg).

How They Work

To calculate the probability of X falling within a specific range [a, b], you need to integrate the PDF from a to b:

$$P(a \leq X \leq b) = \int_a^b f(x)\,dx$$

What is the share of population with weight between 65kg and 75kg?

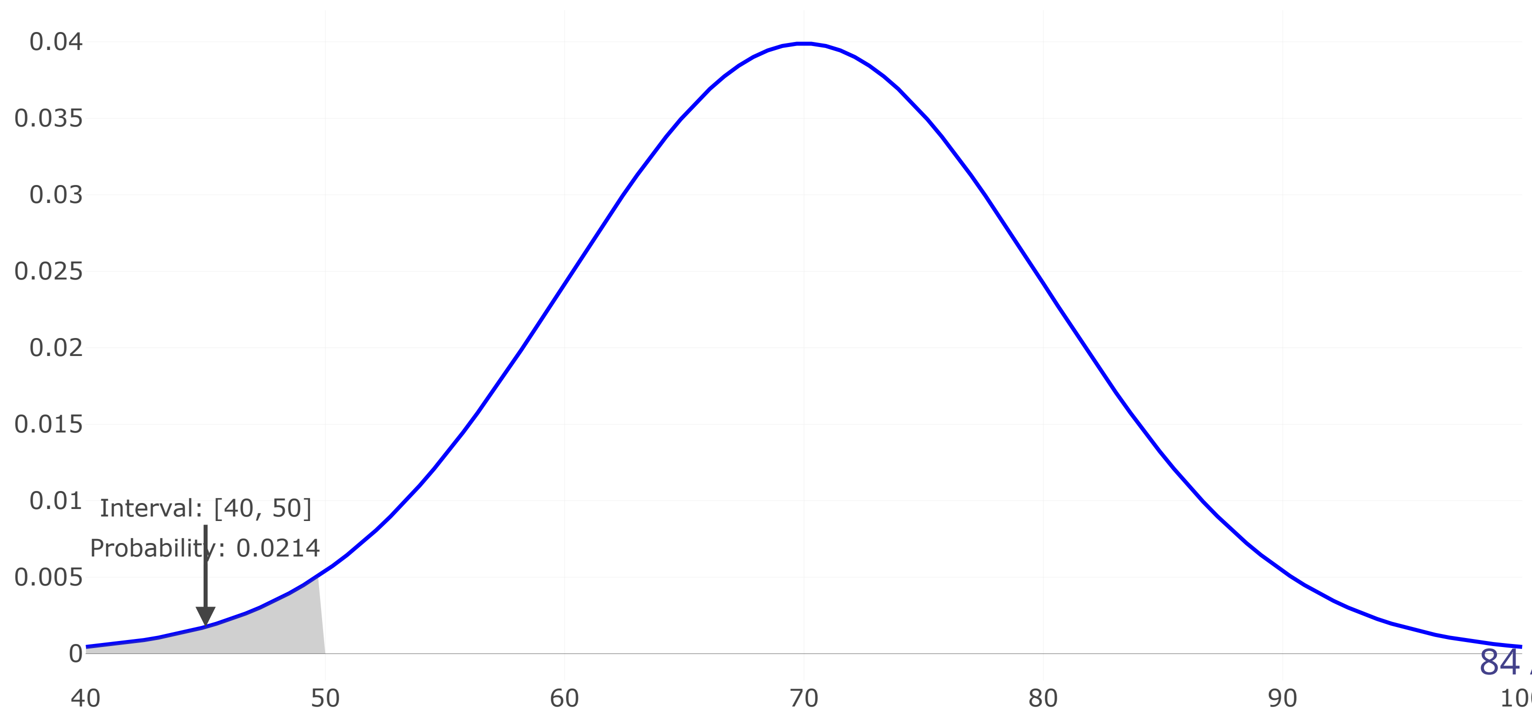


How They Work

To calculate the probability of X falling within a specific range [a, b], you need to integrate the PDF from a to b:

$$P(a \leq X \leq b) = \int_a^b f(x)\, dx$$

What is the share of population with weight between 40 and 50kg?

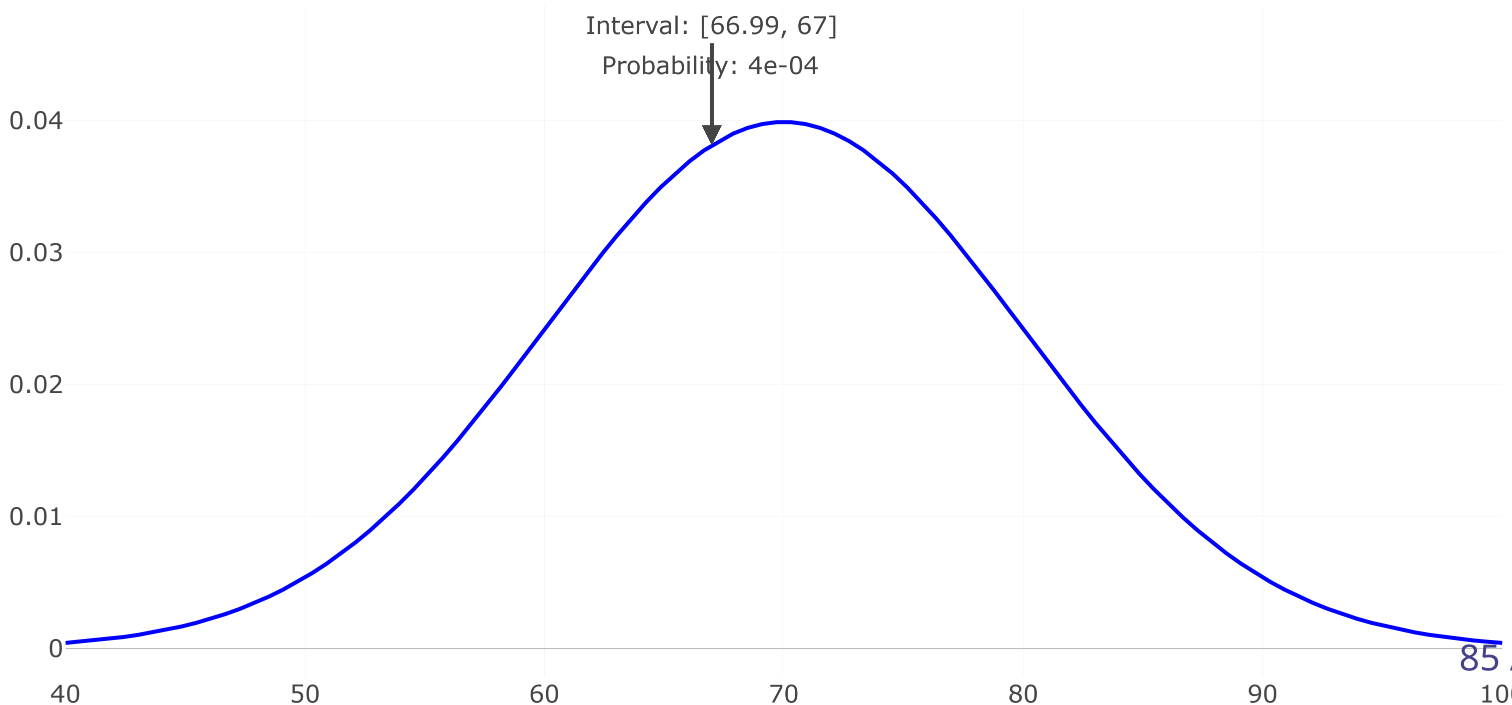


How They Work

To calculate the probability of X falling within a specific range [a, b], you need to integrate the PDF from a to b:

$$P(a \leq X \leq b) = \int_a^b f(x)\,dx$$

What is the share of population with weight between 66.99 and 67 kg?

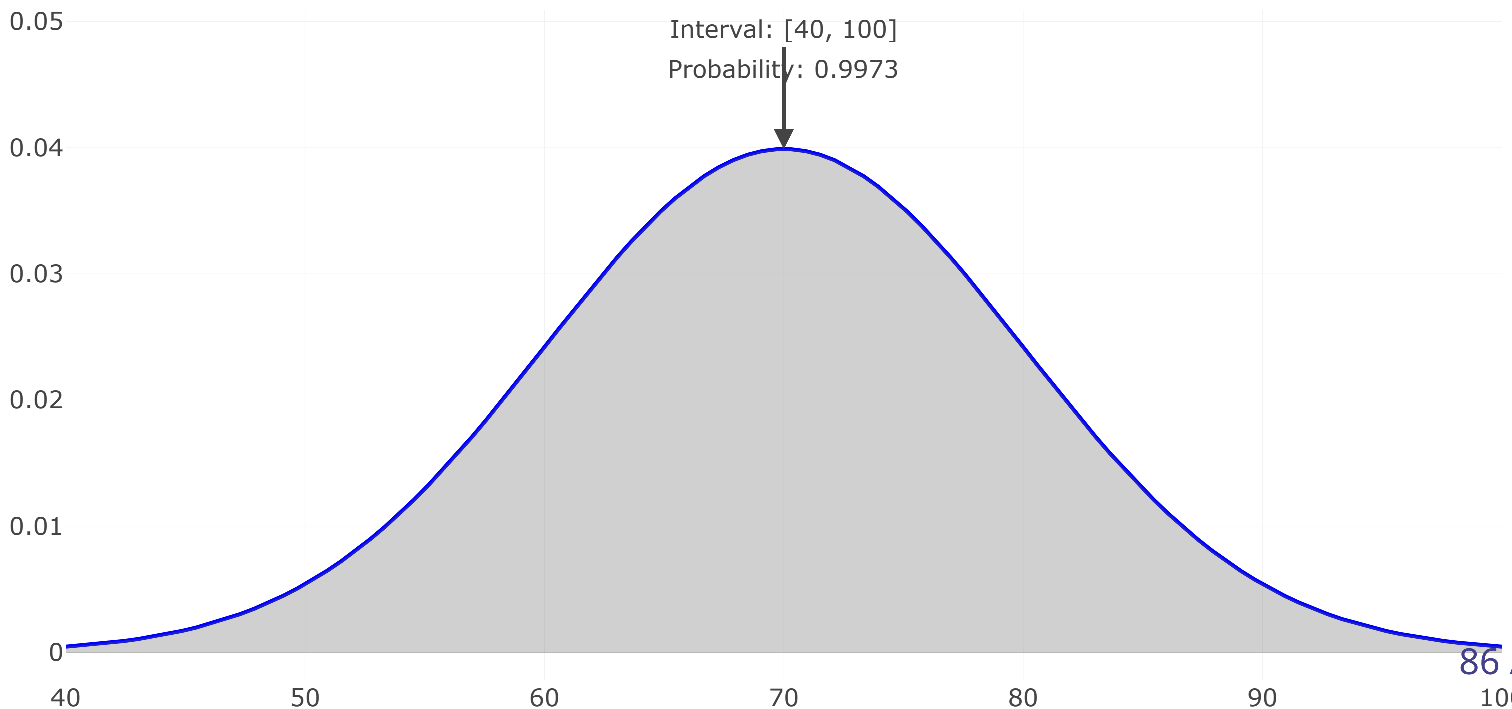


How They Work

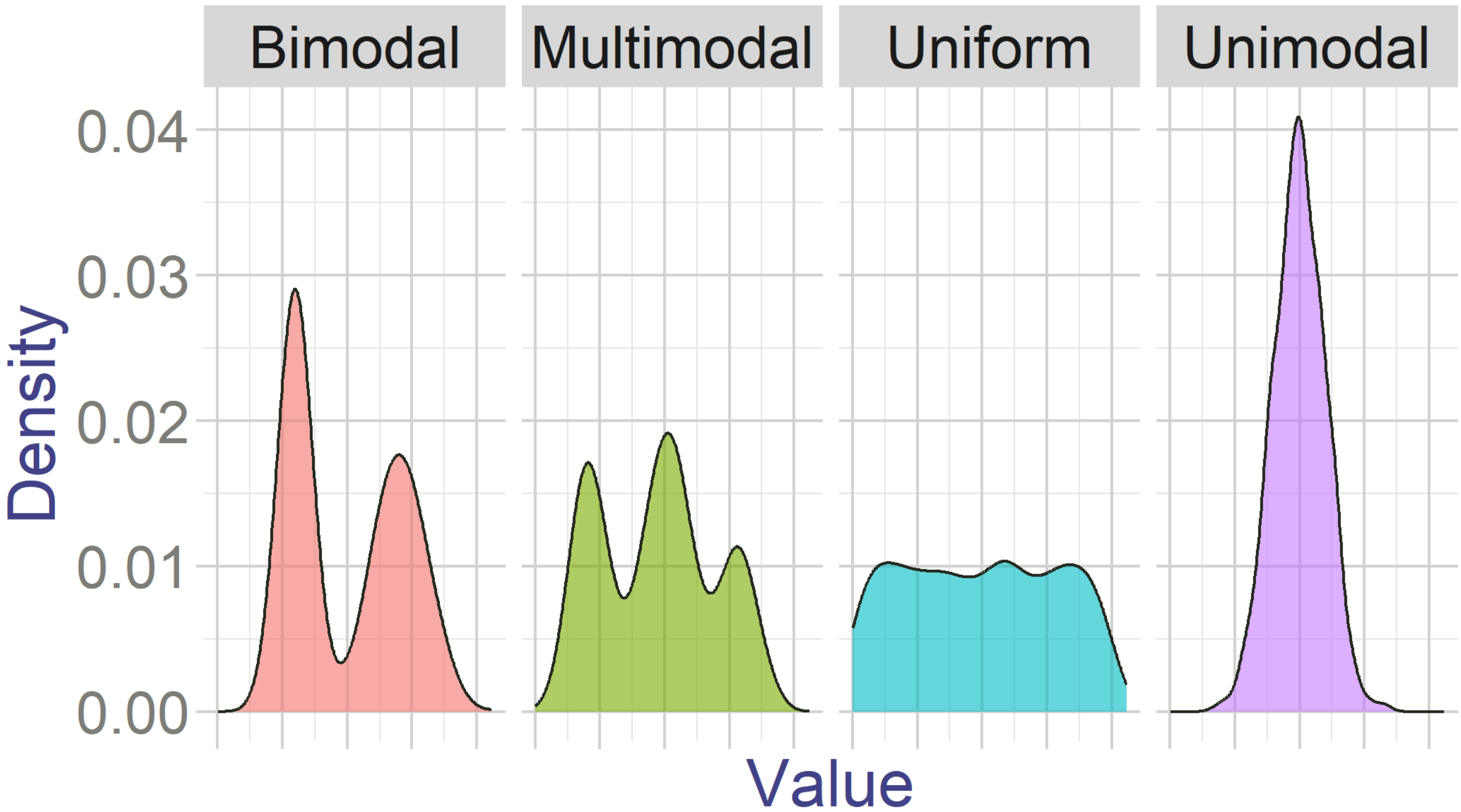
To calculate the probability of X falling within a specific range [a, b], you need to integrate the PDF from a to b:

$$P(a \leq X \leq b) = \int_a^b f(x)\, dx$$

What is the share of population with weight between 40 and 100 kg?



Distribution Shapes: Modality

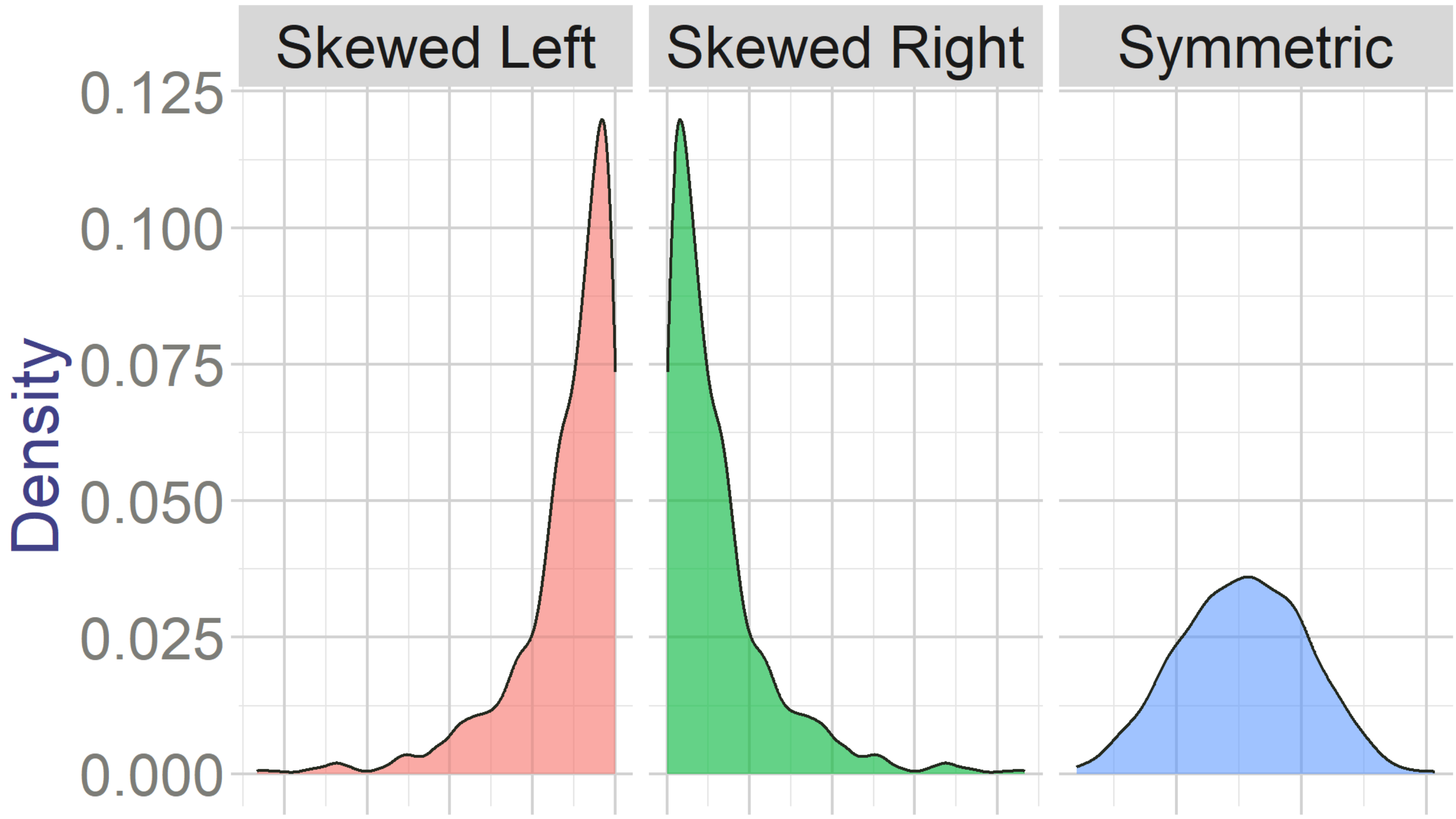


Which is uniformaly distributed

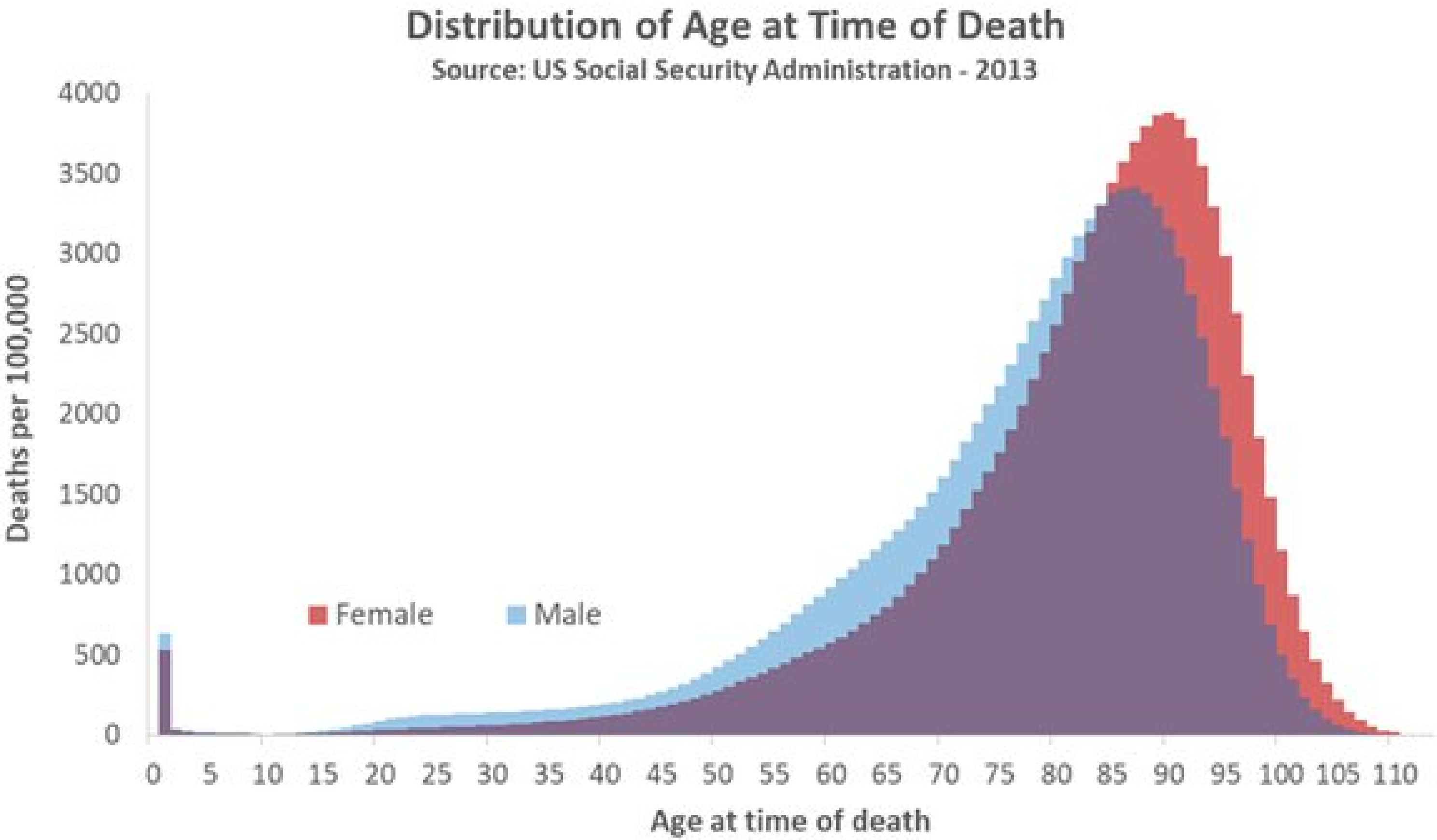
1. Weights of Adult Females
2. Salaries in Mexico
3. Airbnb prices in CDMX
4. Birthdays of Classmates (day of the month)


1 500
1 000
500
0
0
1 000
2 000
3 000
200 000
100 000
0
0
100
200
300
3 000
2 000
1 000
0
50
100
150
200
count
300 000
200 000
100 000
0
0
5 000
10 000
15 000
Value

Distribution Shapes: Skewness



Age at death


Distribution of Age at Time of Death
Source: US Social Security Administration - 2013
Deaths per 100,000
4000
3500
3000
2500
2000
1500
1000
500
0
Female
Male
0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100 105 110
Age at time of death

What if we want to calculate proportion of people who weight less or equal to 50kg?

Cumulative Distribution Function (CDF)

The Cumulative Distribution Function (CDF) gives the probability that a random variable X will take on a value less than or equal to a specific value.

For a continuous random variable X with PDF f(x), the CDF F(x) is defined as:

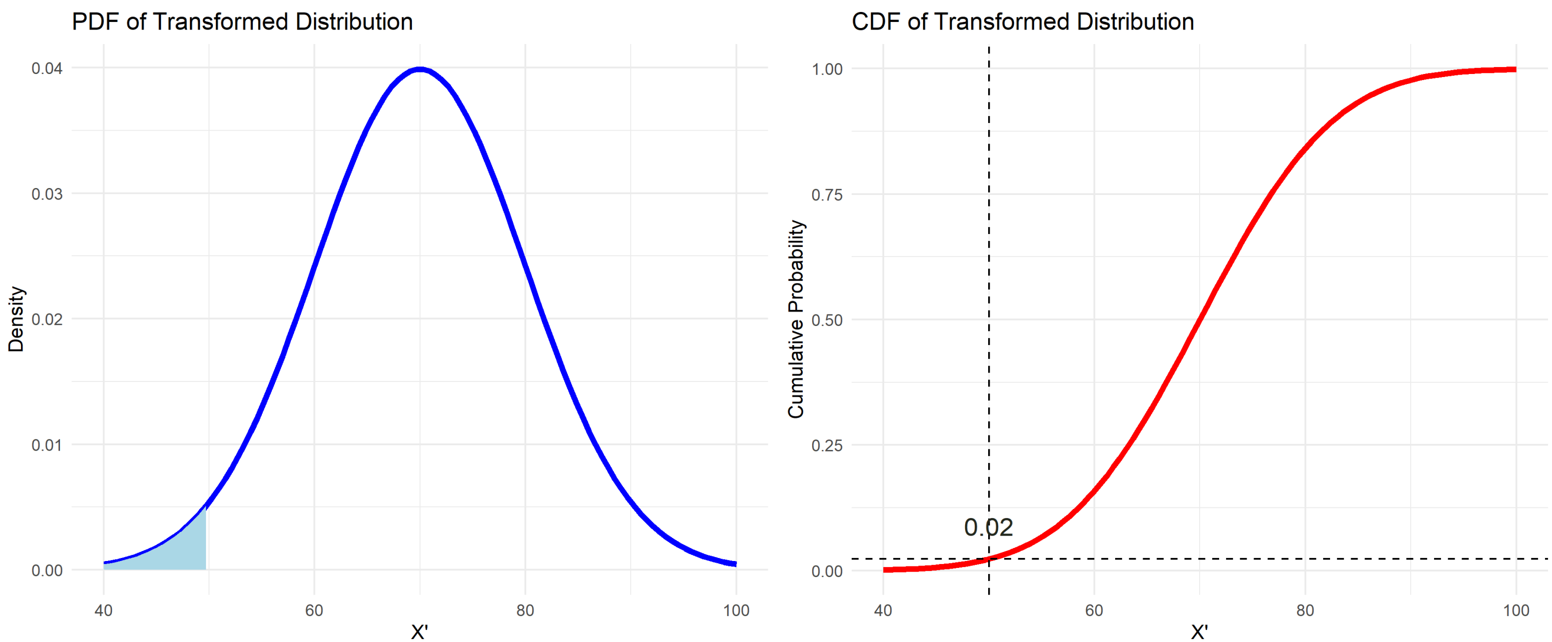
$$F(x) = \int_{-\infty}^{x} f(t)\,dt = P(X \leq x)$$

Characteristics:

- The CDF starts (for minus infinity) at 0 (minimum)
- It approaches 1 as x approaches infinity (maximum)
- It is non decreasing
- It is right continuous

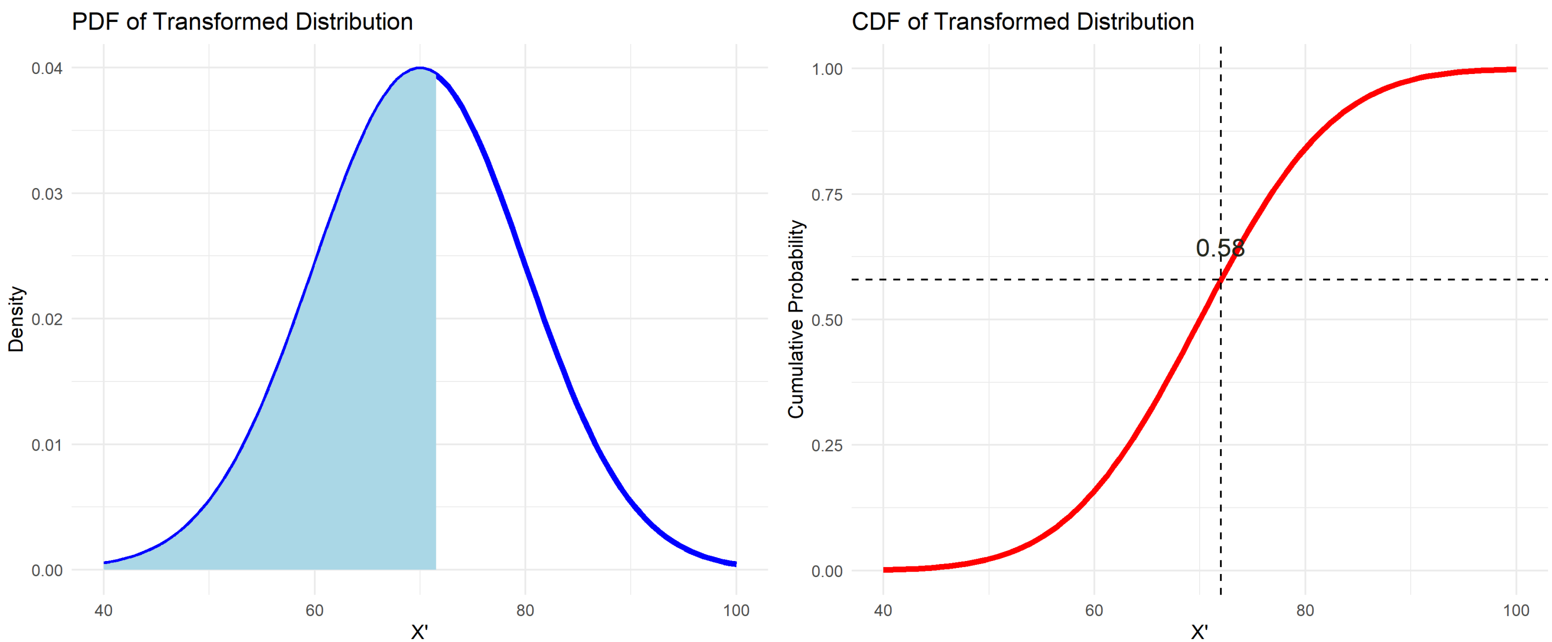
Example 1: Normal Variable (weight in the population)

$$F(50) = \int_{-\infty}^{50} f(t)\,dt = P(X \leq 50) = 0.02$$



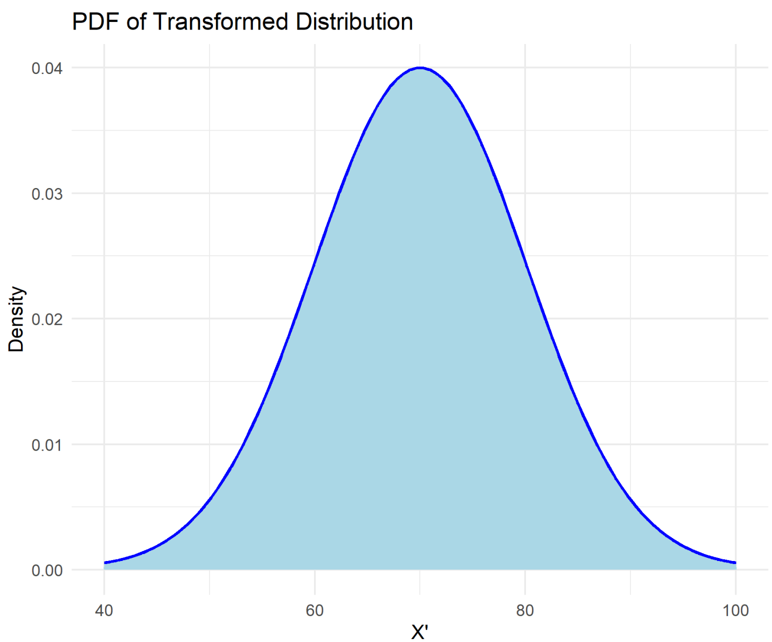
Example 2: Normal Variable (weight in the population)

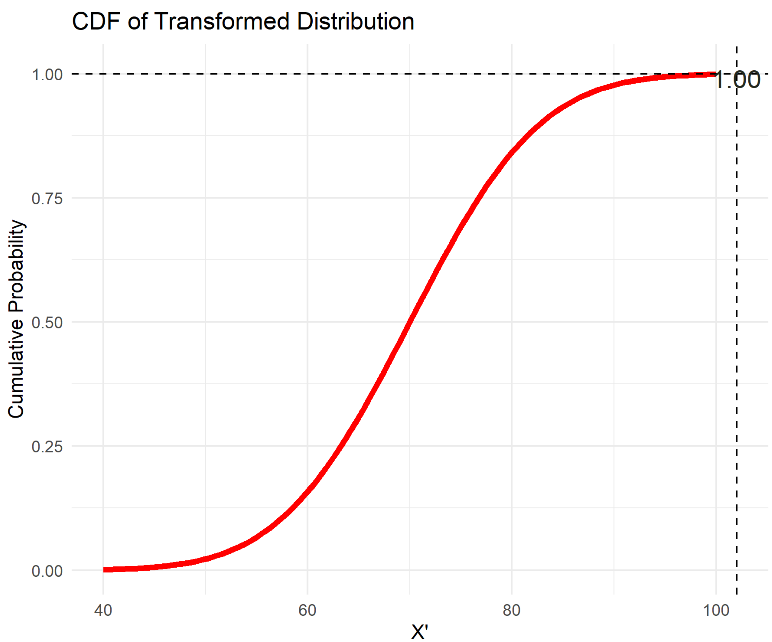
$F(72) = \int_{-\infty}^{72} f(t)\,dt = P(X \leq 72) = 0.58$



Example 3: Normal Variable (weight in the population)

$$F(102) = \int_{-\infty}^{102} f(t)\, dt = P(X \leq 102) = 0.99$$



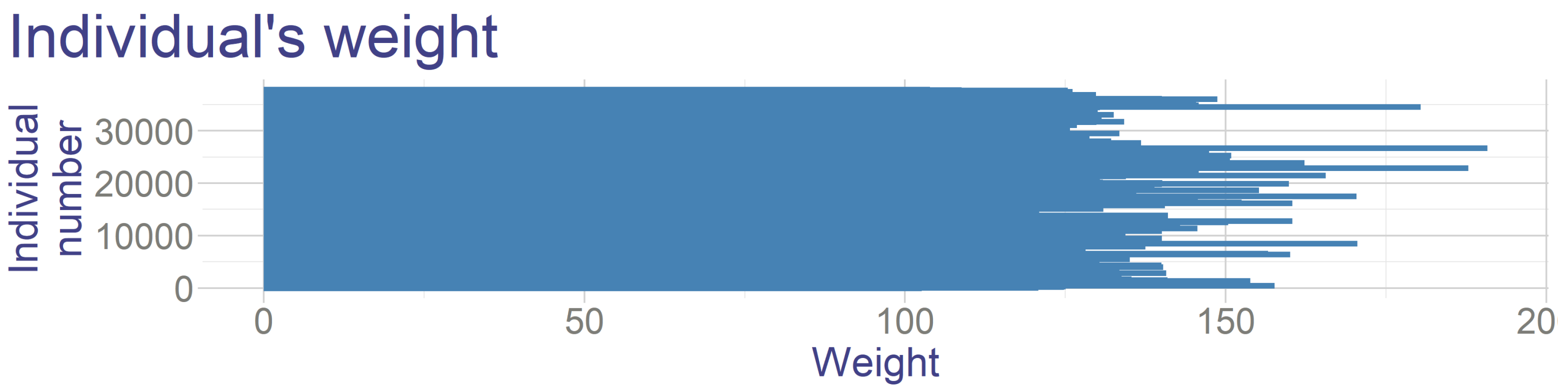


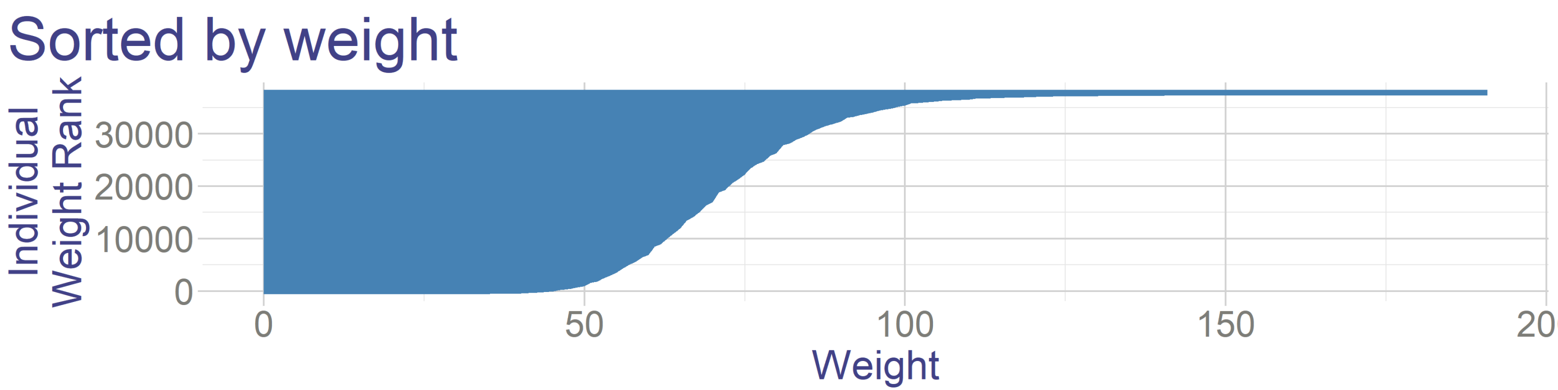
Never integrate a CDF!

Empirical CDF

What if we only have a sample and we don't know the true pdf?

Intuition on how it comes up:

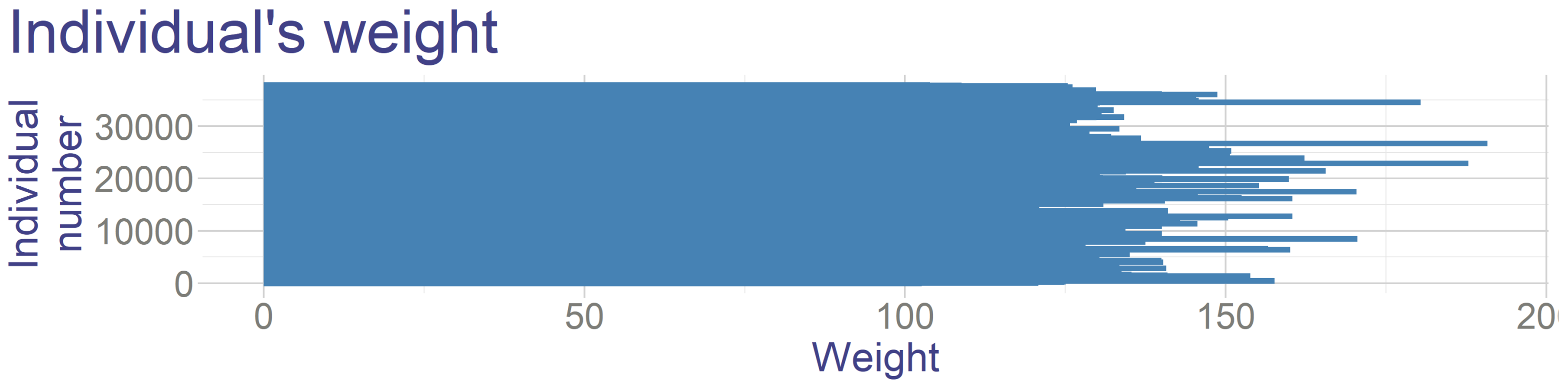


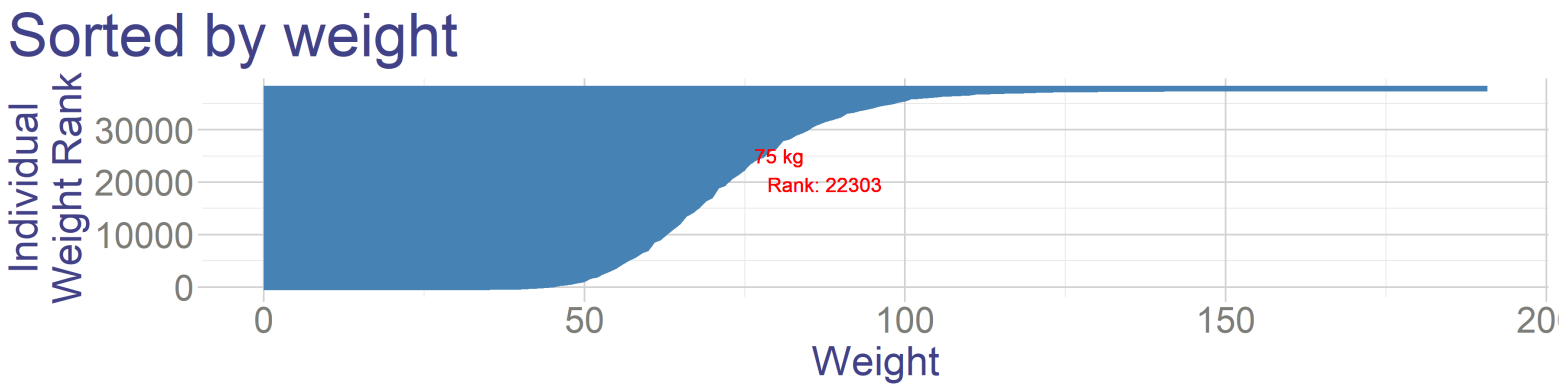


Empirical CDF

What if we only have a sample and we don't know the true pdf?

Intuition on how it comes up:





Empirical CDF

$$ECDF(x) = \frac{\sum I(w_i \leq x)}{N} = \frac{\text{Number of people with weight lower than x}}{N}$$

- $I(w_i < x) = 1$ if weight of person i is lower than x (*Indicator Function*)
- N is total number of people (*Sample Size*)
- Share of people with weight lower than x

- So how do we calculate share of people with weight=<50kg? $P(weight \leq 50) = ECDF(50)$
- What about more than 100?

[6 puntos] Consider two variables that represent the amount of money people spend on rent in Chelsea (X) and Kensignton (Y) in London, measured in Pounds (£), with cummulative density functions $F_X(x)$ and $F_y(y)$ given respectively by

$$F_X(x) = \frac{\sqrt{x - 5,000}}{100}\ \mathbb{I}_{[5,000,15,000)}(x) + \mathbb{I}_{[15,000,\infty)}(x)$$

and

$$F_Y(y) = \frac{y - 4,000}{8,000}\ \mathbb{I}_{[4,000,12,000)}(y) + \mathbb{I}_{[12,000,\infty)}(y)$$

- Is $F_x(X)$ a valid distribution function?
- What's the probability that the rent is larger than 10 000?

Exercises:

- Review Exercises:
  - PDF 1: 1,2,3,4,5,6,7
  - PDF 2: 14,15,16,17