

Exploratory data analysis

Exploring numerical data

Prof. Dr. Jan Kirenz
HdM Stuttgart

Radius is now LendingClub.

[Visit our Banking website](#)

Personal Loans Up to \$40,000

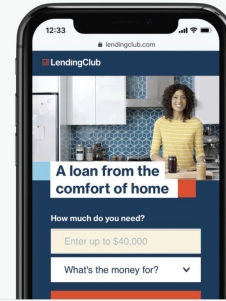
Check your rate. It won't impact your credit score.



 ▾

[Check Your Rate](#)

✉ [Respond to Mail Offer](#)



How LendingClub Works

Apply in Minutes

Get customized loan options based on what you tell us.

Choose a Loan Offer

Select the rate, term, and payment options you like best.

Get Funded

Once your loan is funded, we'll send the money straight to your bank account or pay your creditors directly.

[Check Your Rate](#)

Join Over 3 Million Members



Thank you so much for valuing me as a customer, and coming through for me and my family at a trying time in this world.

— Roselyn, a member from Texas

[Read More Reviews](#)


\$60 Billion+
Borrowed



3 Million+
Members



★★★★★
54,898 Reviews



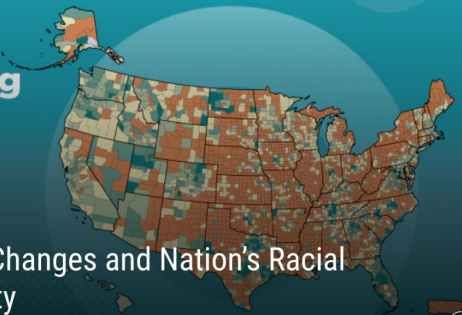
Data for 3142 counties in the United States



Search

[BROWSE BY TOPIC](#)[EXPLORE DATA](#)[LIBRARY](#)[SURVEYS/ PROGRAMS](#)[INFORMATION FOR...](#)[FIND A CODE](#)[ABOUT US](#)

2020 Census Redistricting Data



Local Population Changes and Nation's Racial and Ethnic Diversity

Read More

The U.S. Census Bureau today released additional 2020 Census results showing an increase in the population of U.S. metro areas compared to a decade ago.

SURVEYS

Help for Survey Participants

Verify that the survey you received is real and learn how to respond.

QUICKFACTS

Access Local Data

Learn about your community, county, state and the U.S. It's fast, easy and shareable.

POPULATION CLOCK

August 22, 2021

USA
332,659,369

World
7,784,236,581

U.S. CENSUS BUREAU ECONOMIC INDICATORS

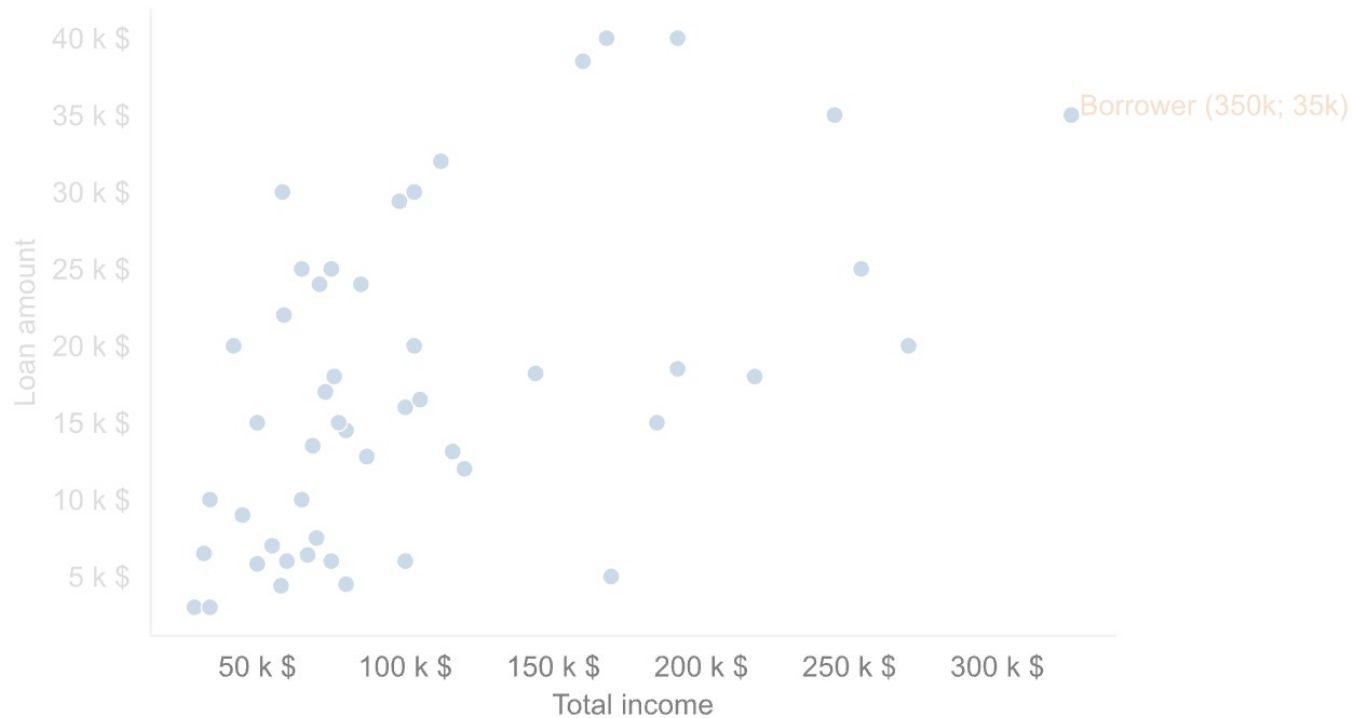
Selected Services Revenue 2nd Quarter 2021 Report Released 10:00 AM EDT, 8/19/21	\$4,367.8 B Advance Report 4.0%
New Residential Construction July 2021 Report Released 8:30 AM EDT, 8/18/21	1,534,000 Housing starts -7.0%
Business Inventories June 2021 Report Released 10:00 AM EDT, 8/17/21	\$2,057.4 B 0.8%
Advance Monthly Retail Sales July 2021 Report Released 8:30 AM EDT, 8/17/21	\$617.7 B -1.1%

[All Economic Indicators](#)

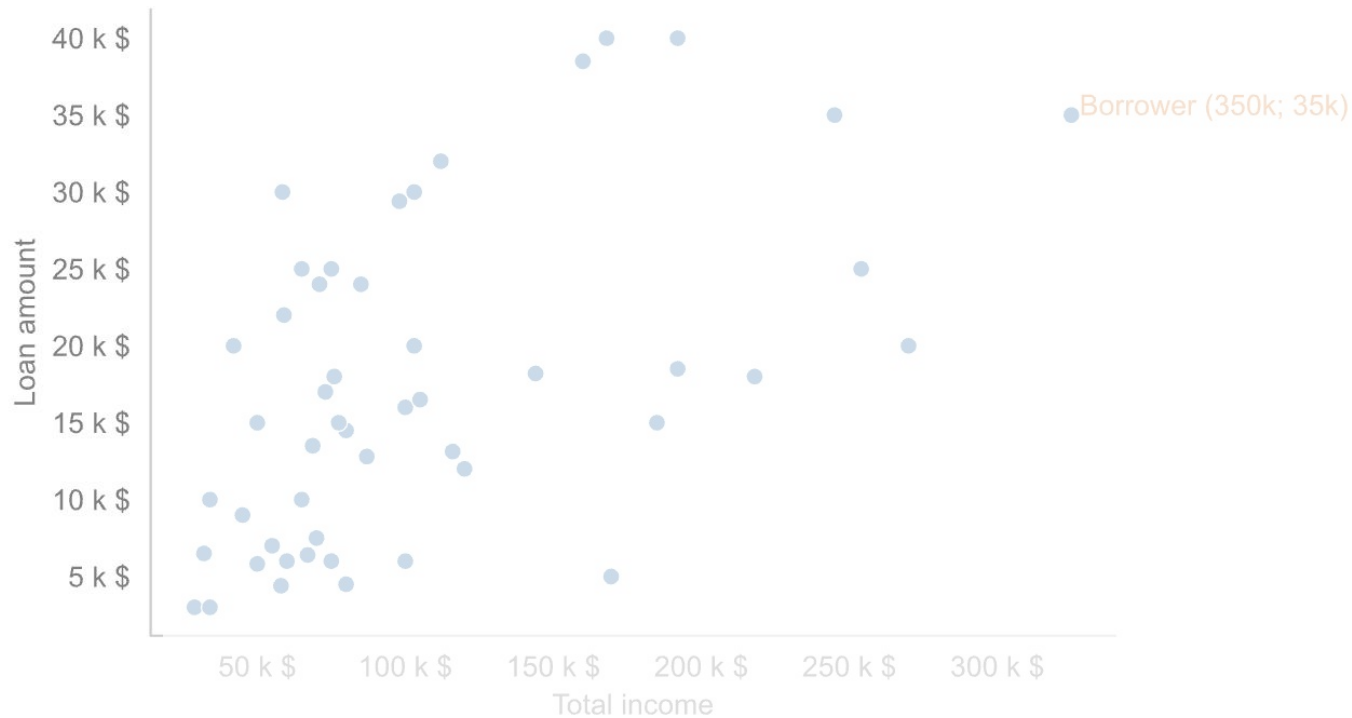
* change not statistically significant
○ significance not reported / applicable

Scatterplots for paired data

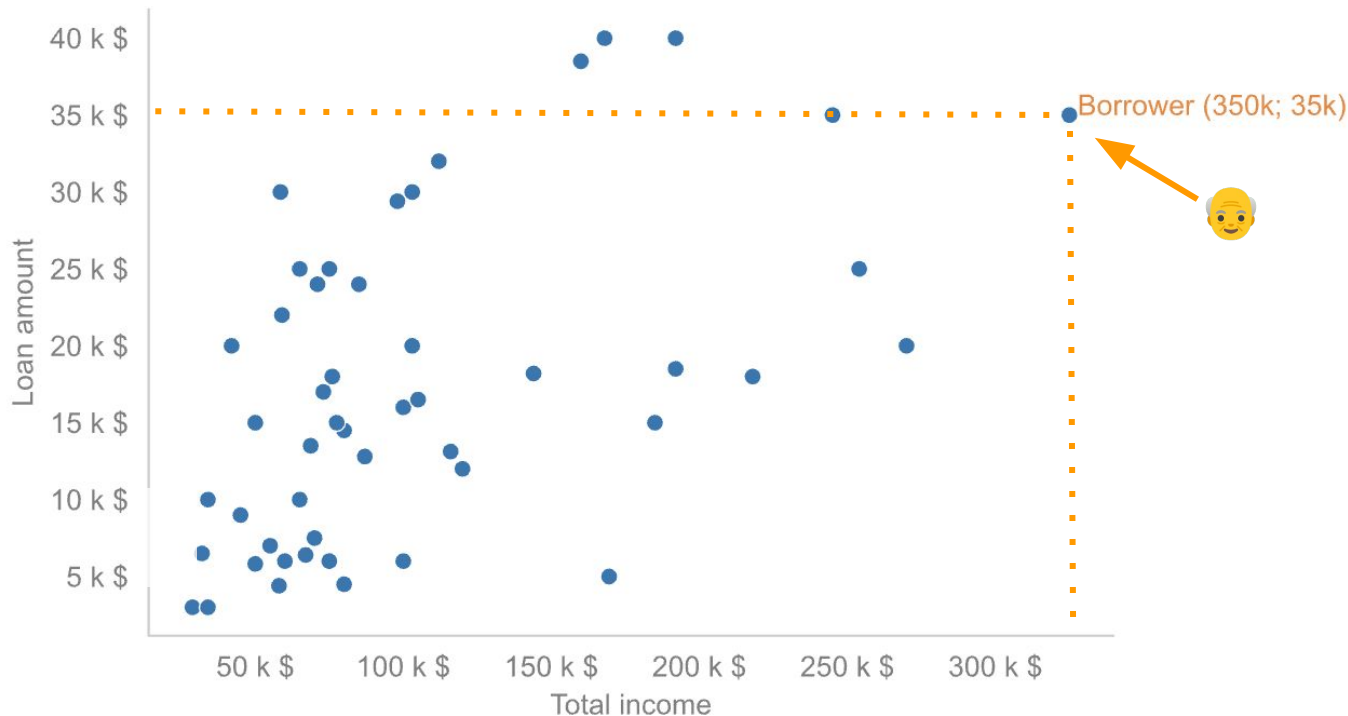
Total income is on the x-axis



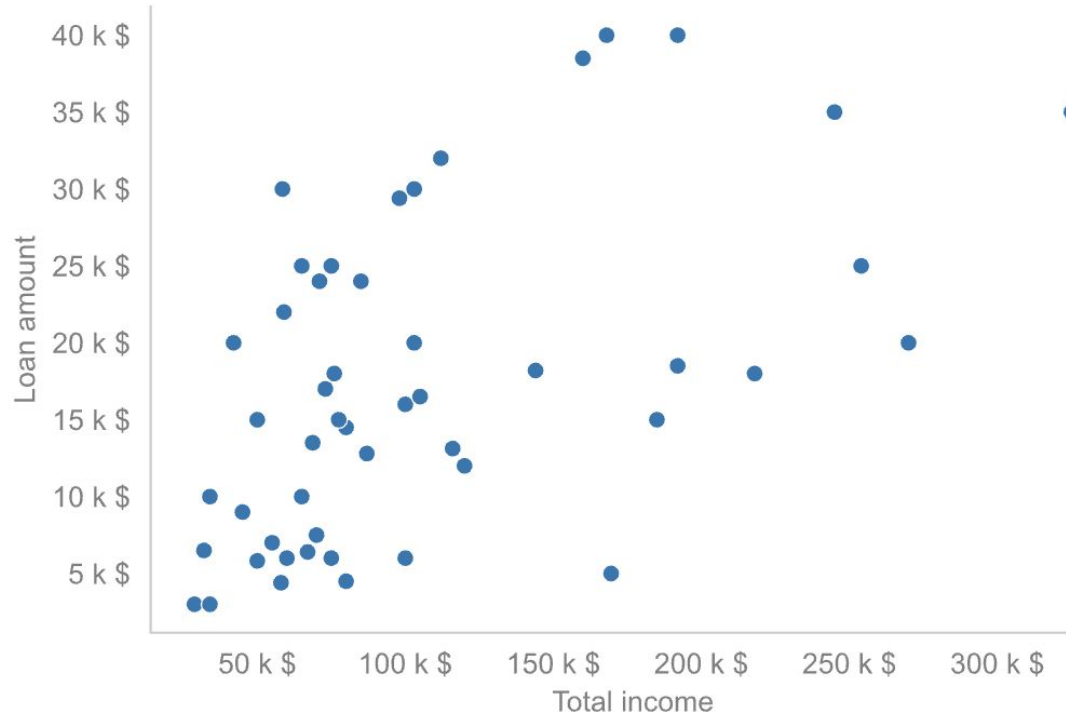
Loan amount is on the y-axis



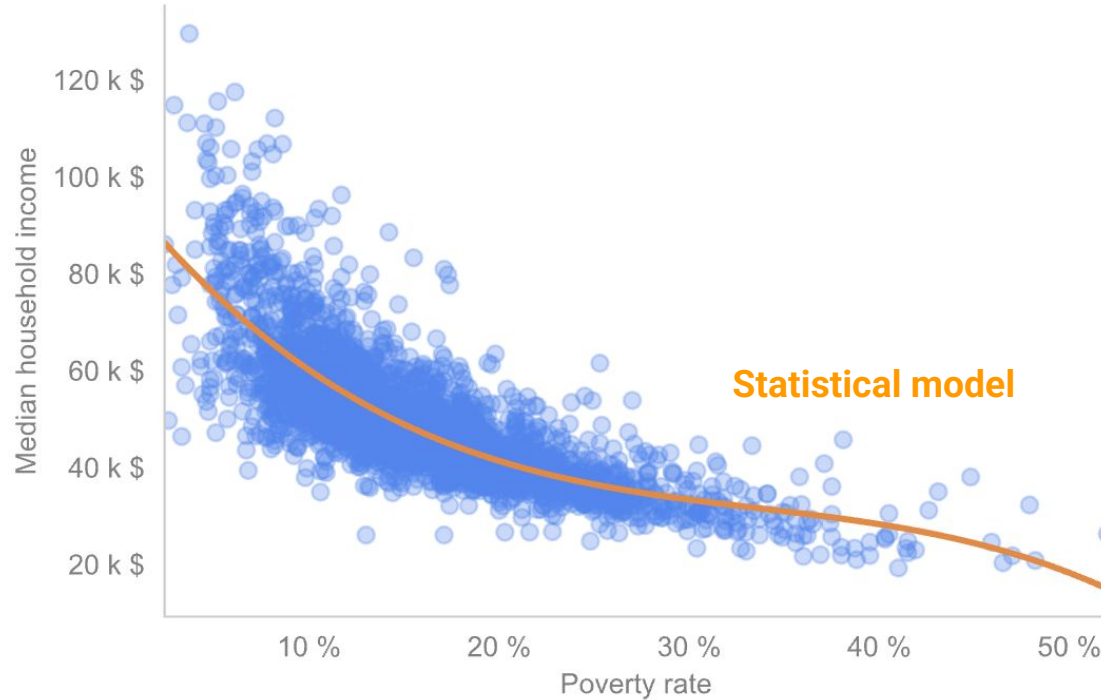
This is a borrower with a **total income** of 350 k and a **loan amount** of 35 k



Each point represents a single case (borrower)

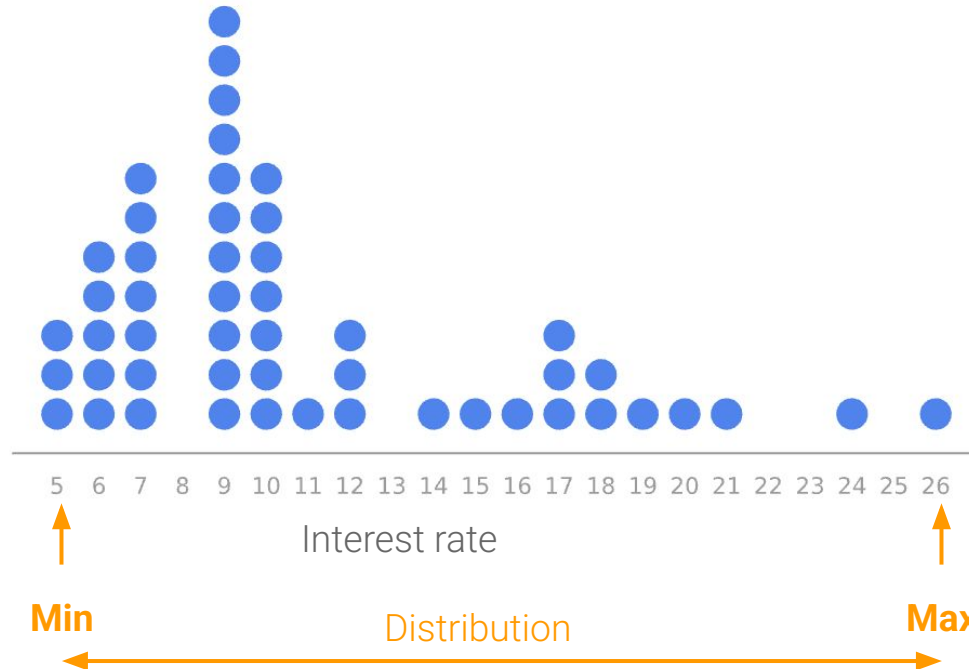


A scatterplot of the median household income against the poverty rate for the county dataset



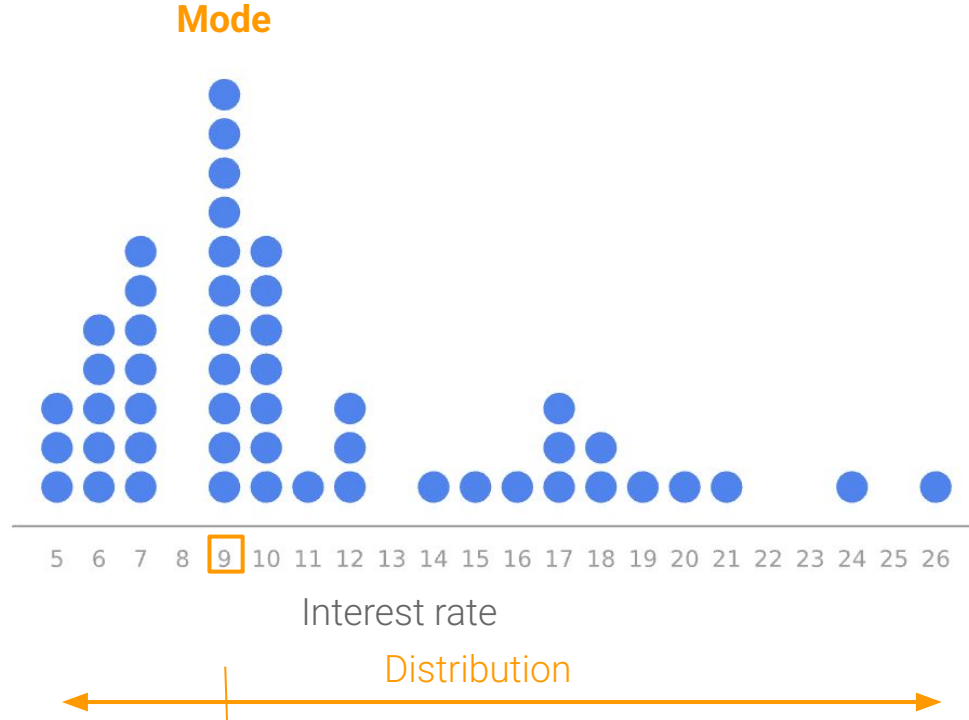
Dot plots

A dot plot of interest rate (numbers are rounded)



$n = 50$ (there are 50 dots)

The **mode** is the value with the most occurrences



Median

- If the data are ordered from smallest to largest, the median is the observation right in the **middle**.
- If there are an **even number** of observations, there will be two values in the middle, and the median is taken as their average.

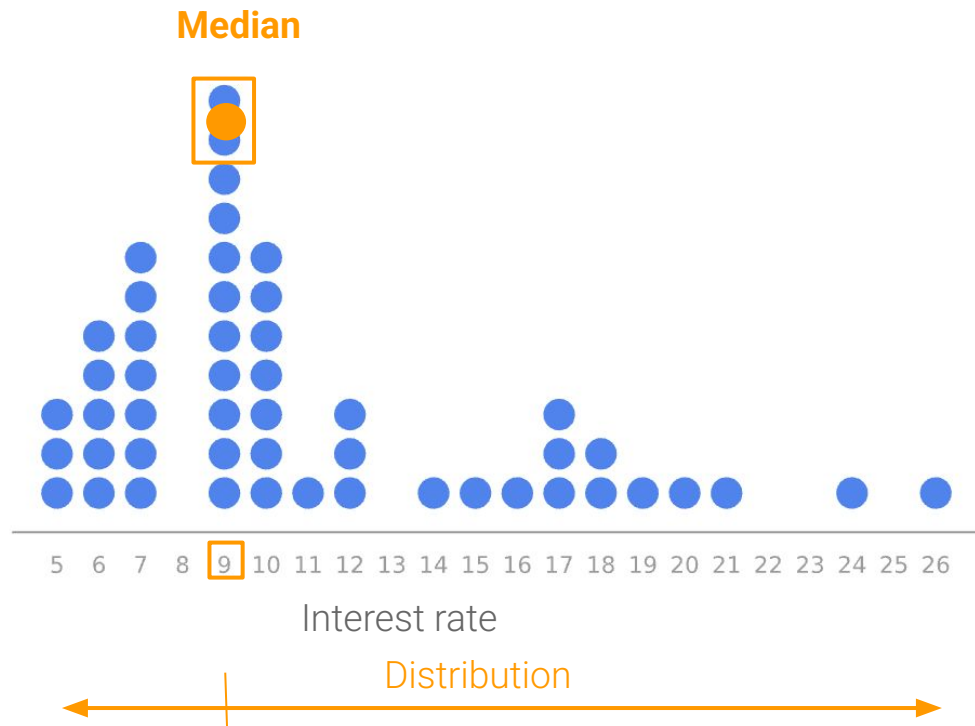
Interest rates from the loan50 dataset, arranged in ascending order (original values)

	1	2	3	4	5	6	7	8	9	10
1	5.31	5.31	5.32	6.08	6.08	6.08	6.71	6.71	7.34	7.35
10	7.35	7.96	7.96	7.96	7.97	9.43	9.43	9.44	9.44	9.44
20	9.92	9.92	9.92	9.92	9.93	9.93	10.42	10.42	10.90	10.90
30	10.91	10.91	10.91	11.98	12.62	12.62	12.62	14.08	15.04	16.02
40	17.09	17.09	17.09	18.06	18.45	19.42	20.00	21.45	24.85	26.30

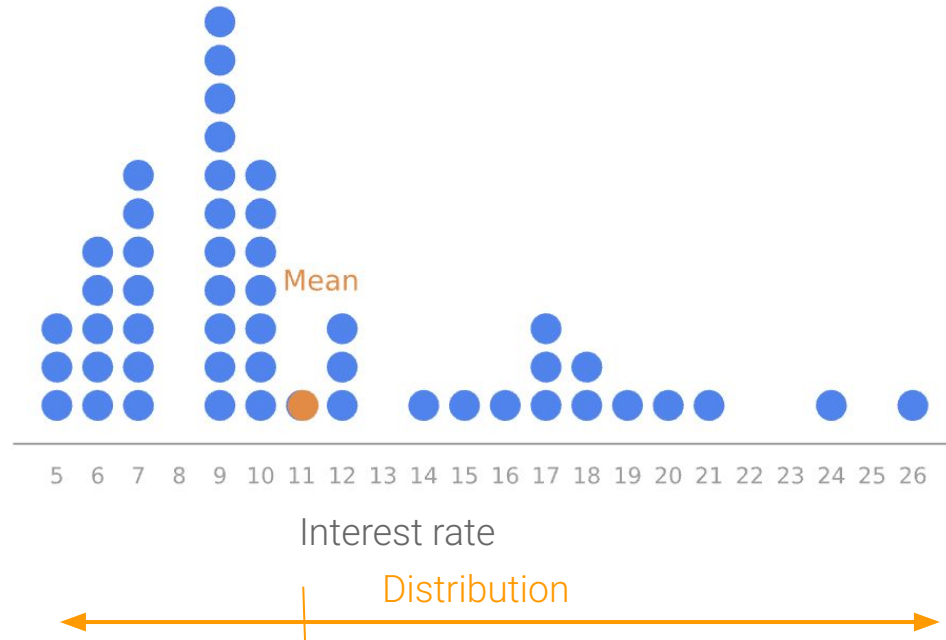
The middle

$$\text{Median} = (9.93 + 9.93) / 2 = 9.93$$

The **median** is the number in the middle



The **mean** is the center of the distribution (the average)



The **sample mean**

The sample mean, denoted as $\bar{x}$, can be calculated as

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n},$$

where $x_1, x_2, \dots, x_n$ represent the n observed values.

- The sample mean is a sample statistic, and serves as a **point estimate** of the population mean.
- This estimate may not be perfect, but if the sample is good (**representative** of the population), it is usually a pretty good estimate.

Results of a trial of 1500 adults that suffer from asthma

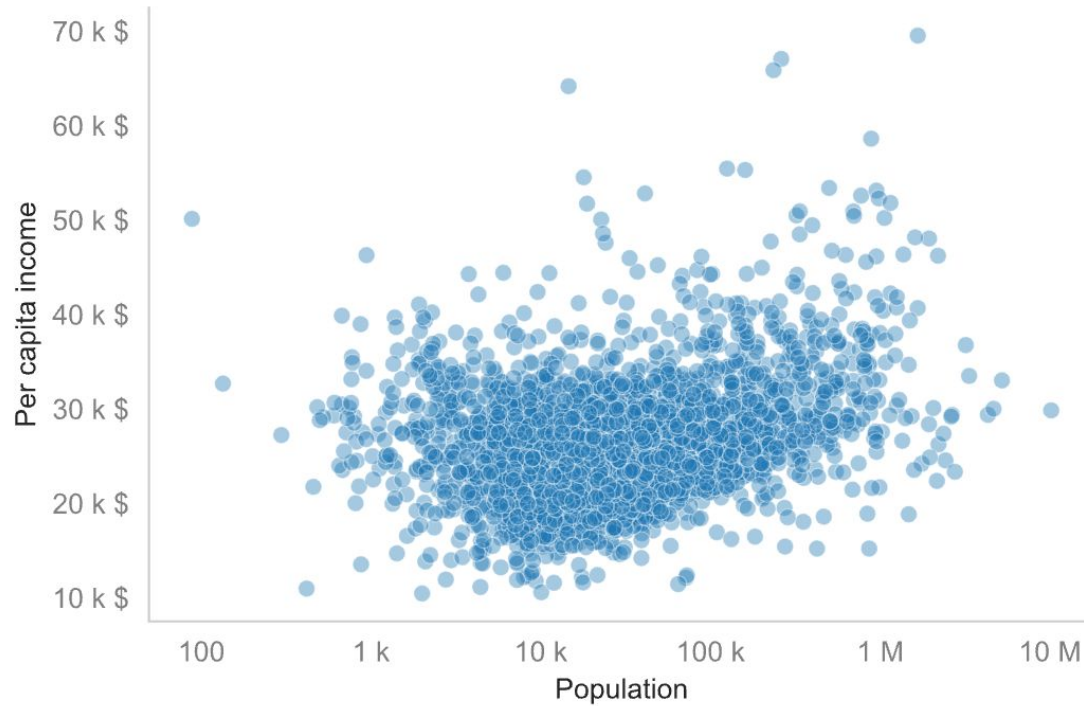
	Description	New drug	Standard drug
0	Number of patients	500	1000
1	Total asthma attacks	200	300

- New drug: $200/500 = 0.4$ asthma attacks per patient
- Standard drug: $300/1000 = 0.3$ asthma attacks per patient

The **population mean**

- The **population mean** is also computed the same way but is denoted as μ .
- It is often **not possible** to calculate μ since population data are rarely available

Per capita income against population size in 3,143 US counties



The **weighted mean**

The weighted mean can be calculated as

$$\text{weighted mean of } x_i\text{'s} = \frac{w_1x_1 + w_2x_2 + \cdots + w_nx_n}{w_1 + w_2 + \cdots + w_n} = \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i}$$

$x_1, x_2, \dots, x_n$ represent the n observed values.

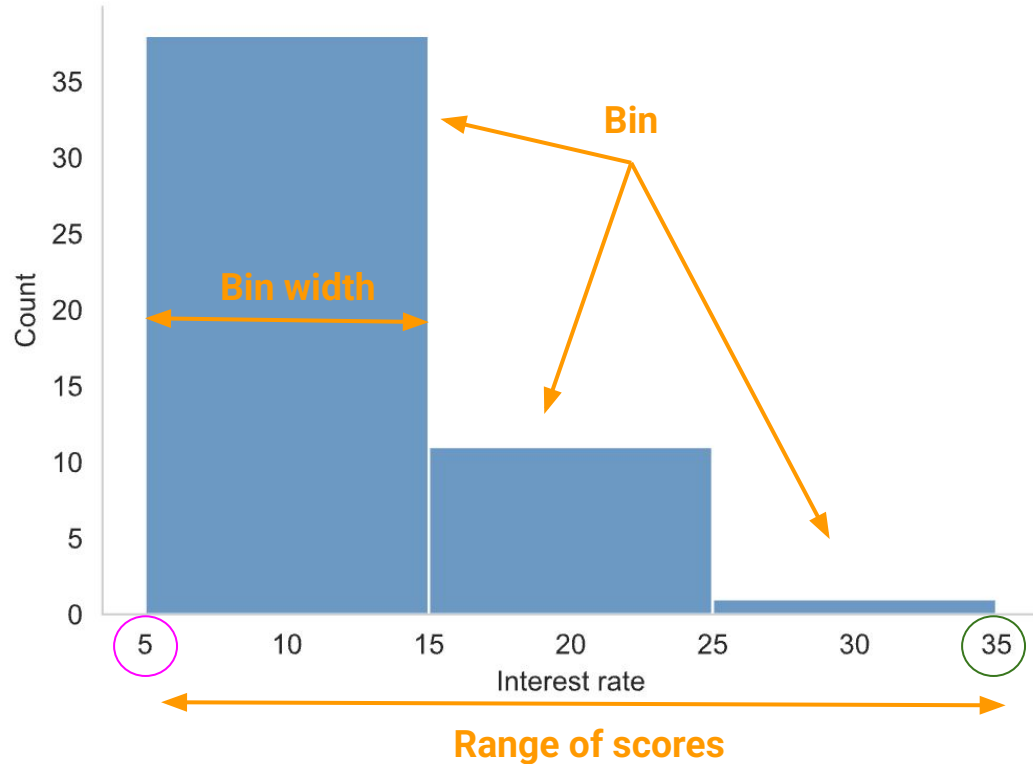
$w_1, w_2, \dots, w_n$ represent the n weights.

- The simple mean is a weighted mean where all the weights are 1:

$$\frac{x_1 + x_2 + x_3 + \cdots + x_n}{n} = \frac{1 \times x_1 + 1 \times x_2 + 1 \times x_3 + \cdots + 1 \times x_n}{1 + 1 + 1 + \cdots + 1}$$

Histograms

A histogram for interest rate with 3 **bins**, a **bin width** of 10 and a **range of scores** of 30



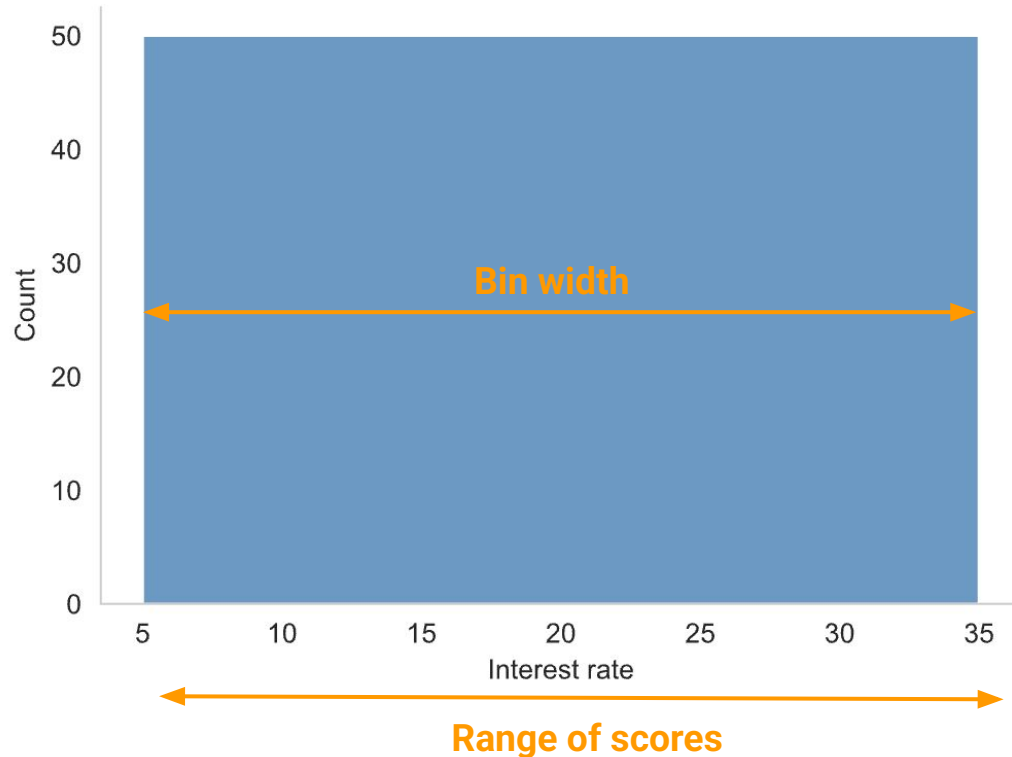
bin width = range of scores / number of bins

$$10 = (35 - 5) / 3$$

number of bins = range of scores / bin width

$$3 = (35 - 5) / 10$$

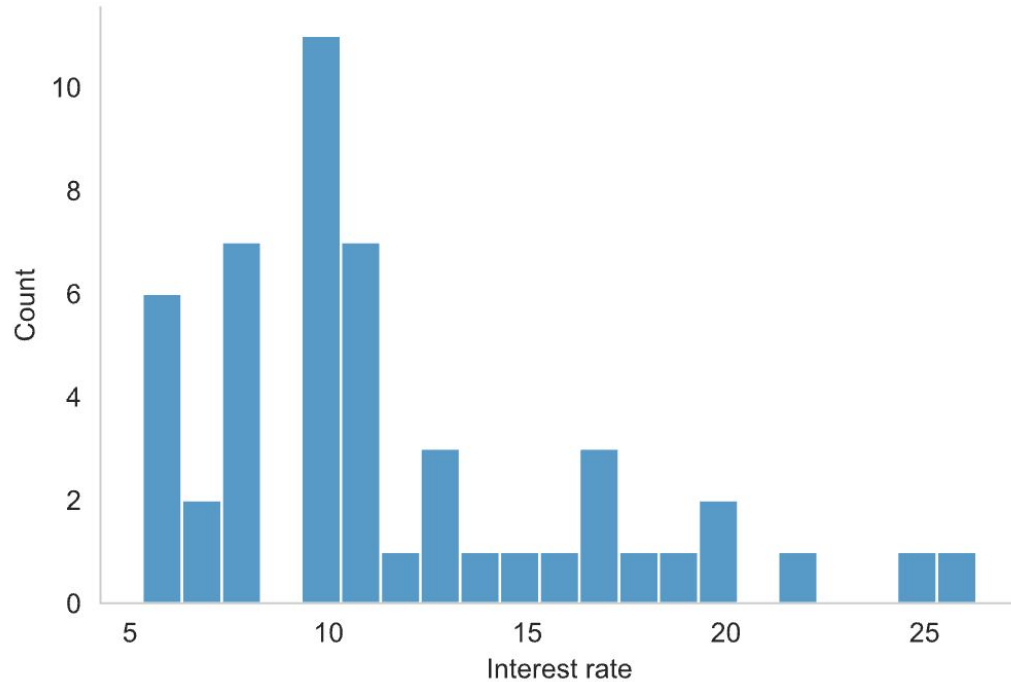
A histogram for interest rate with 1 **bin**, **bin width** of 30 and a **range of scores** of 30



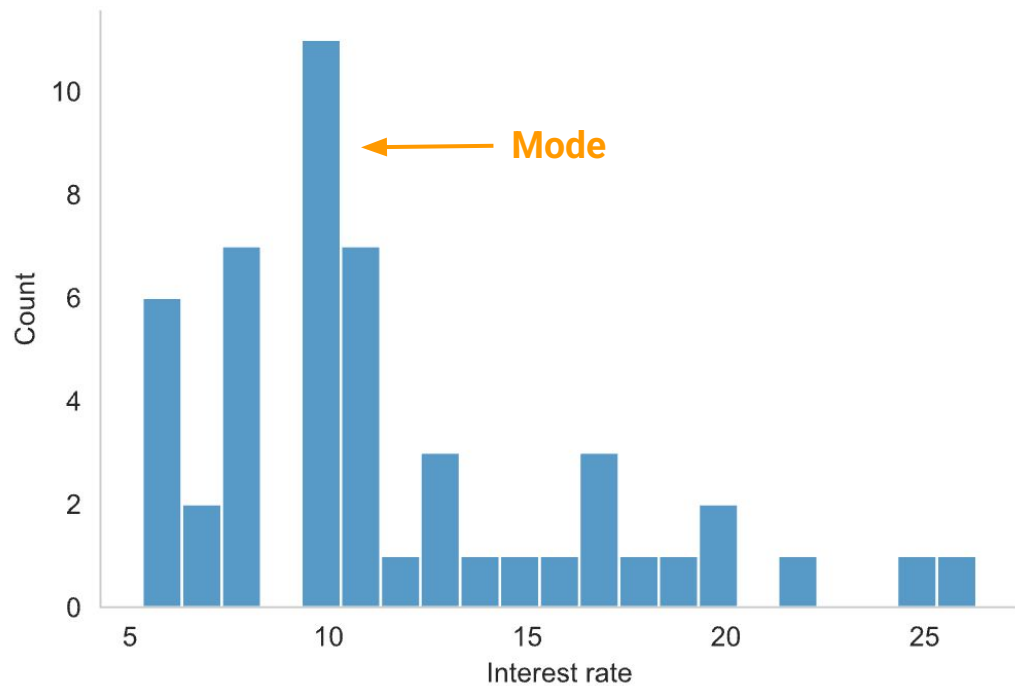
number of bins = range of scores / bin width

$$1 = (35 - 5) / 30$$

A histogram for interest rate with **bin width** 1



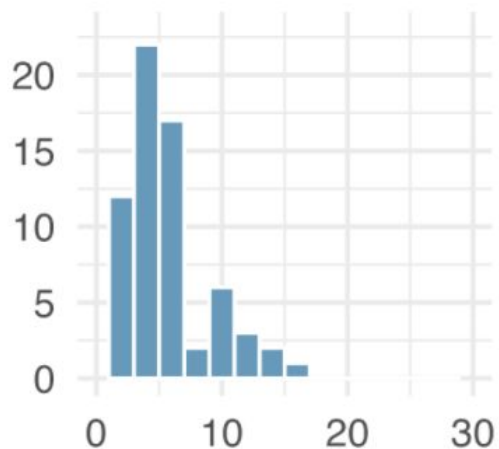
Histograms can be used to identify **modes**



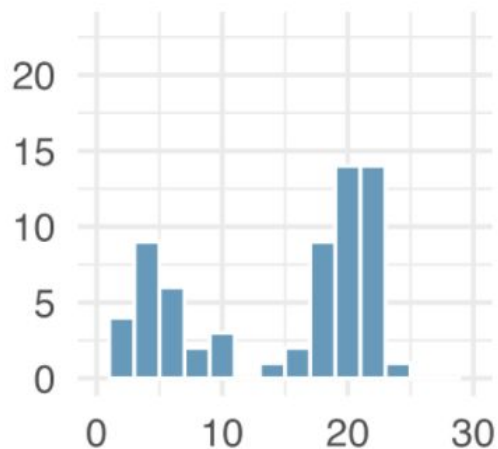
- A mode is represented by a prominent peak in the distribution.

Counting only prominent peaks.

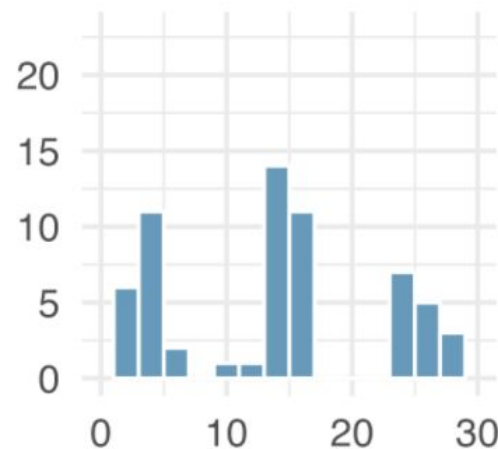
Unimodal



Bimodal



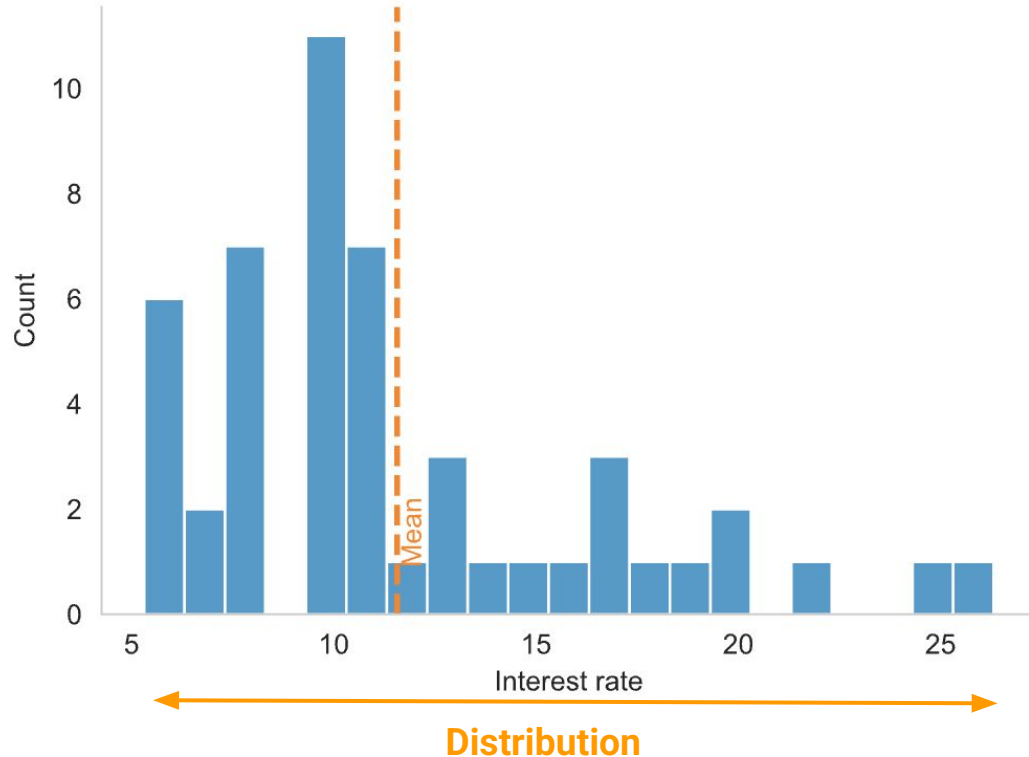
Multimodal



Variance and standard deviation

Example in Google Sheets

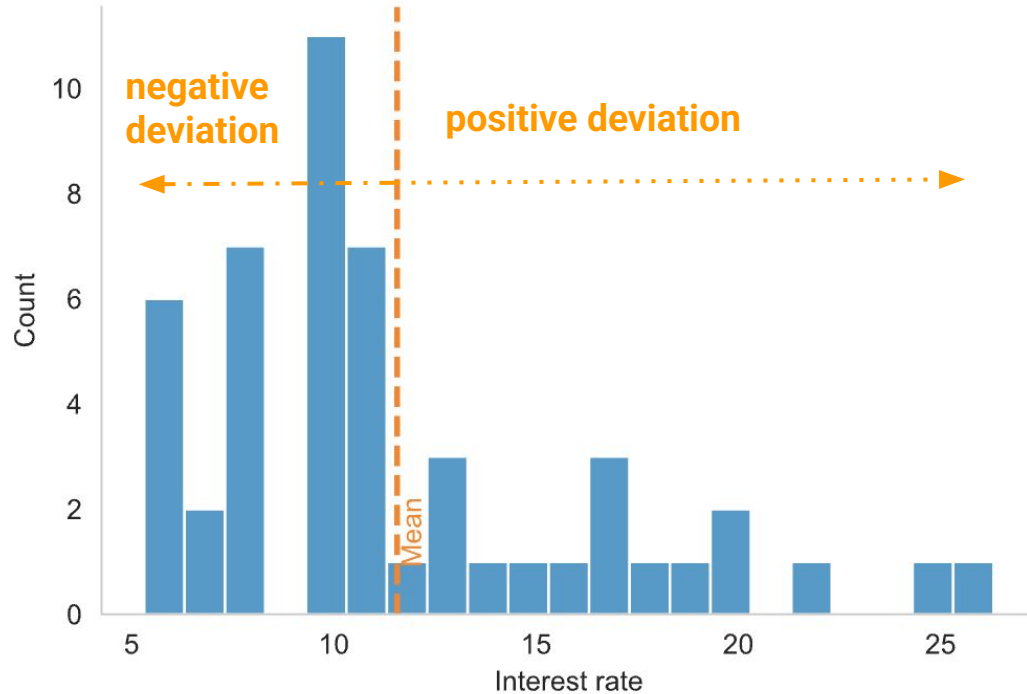
Understanding the shape of the data distribution.



A histogram for interest rate with bin width 1.

- When data trail off in one direction, the distribution has a **long tail**.
- If a distribution has a long left tail, it is **left skewed**.
- If a distribution has a long right tail, it is **right skewed**.

We call the distance of an observation from its mean its **deviation**.



$$\begin{aligned}x_1 - \bar{x} &= 10.9 - 11.57 = -0.67 \\x_2 - \bar{x} &= 9.92 - 11.57 = -1.65 \\x_3 - \bar{x} &= 26.3 - 11.57 = 14.73 \\&\vdots \\x_{50} - \bar{x} &= 6.08 - 11.57 = -5.49\end{aligned}$$

The **sample variance**

Deviation:

$$\begin{aligned}x_1 - \bar{x} &= 10.9 - 11.57 = -0.67 \\x_2 - \bar{x} &= 9.92 - 11.57 = -1.65 \\x_3 - \bar{x} &= 26.3 - 11.57 = 14.73 \\&\vdots \\x_{50} - \bar{x} &= 6.08 - 11.57 = -5.49\end{aligned}$$

$$\begin{aligned}s^2 &= \frac{(-0.67)^2 + (-1.65)^2 + (14.73)^2 + \dots + (-5.49)^2}{50 - 1} \\&= \frac{0.45 + 2.72 + \dots + 30.14}{49} \\&= 25.52\end{aligned}$$

If we square these deviations and then take an average, the result is equal to the **sample variance**, denoted by s^2

Variance & standard deviation

- The **variance** is the average squared distance from the mean.
- The **standard deviation** is the square root of the variance.

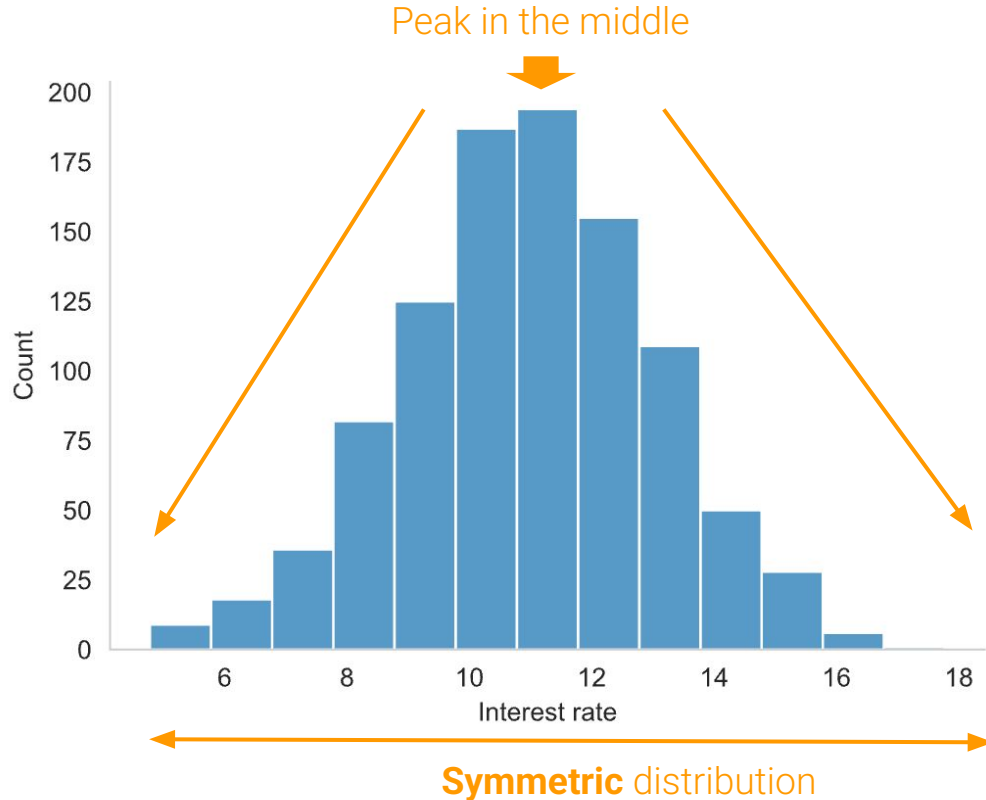
$$s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}}$$

$$s = \sqrt{25.52} = 5.05$$

Standard deviation

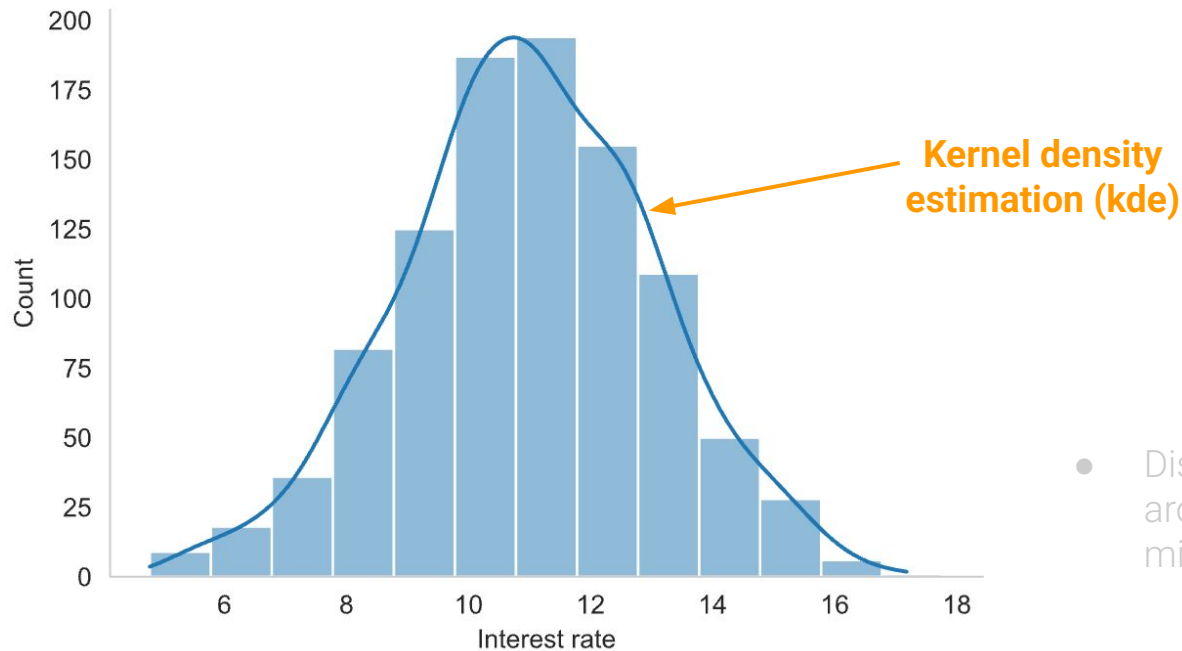
- The standard deviation is useful when considering how far the data are **distributed** from the mean.
- The standard deviation represents the **typical deviation** of observations from the mean.
- Often about **68%** of the data will be within **one standard deviation** of the mean
- and about **95%** will be within **two standard deviations**.
- However, these percentages are not strict rules.

Synthetic generated data with symmetric distribution



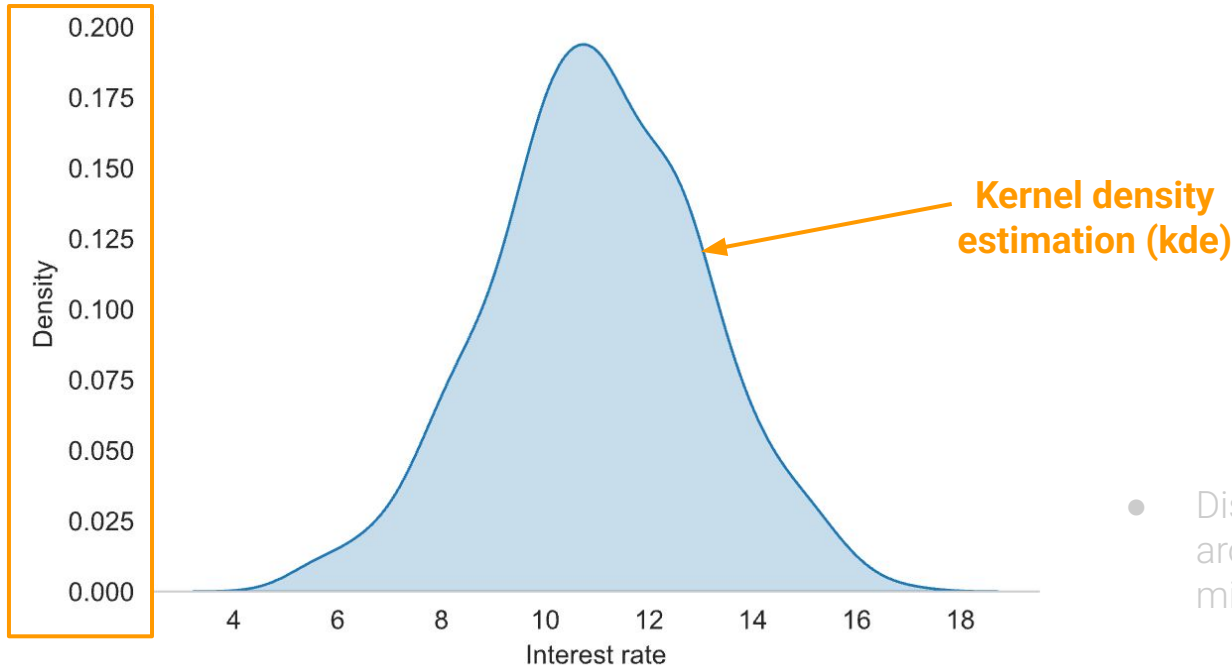
- Variables that show roughly equal trailing off in both directions are called **symmetric**.
- Normal (or Gaussian) distribution

Histogram with kernel density estimation (KDE)



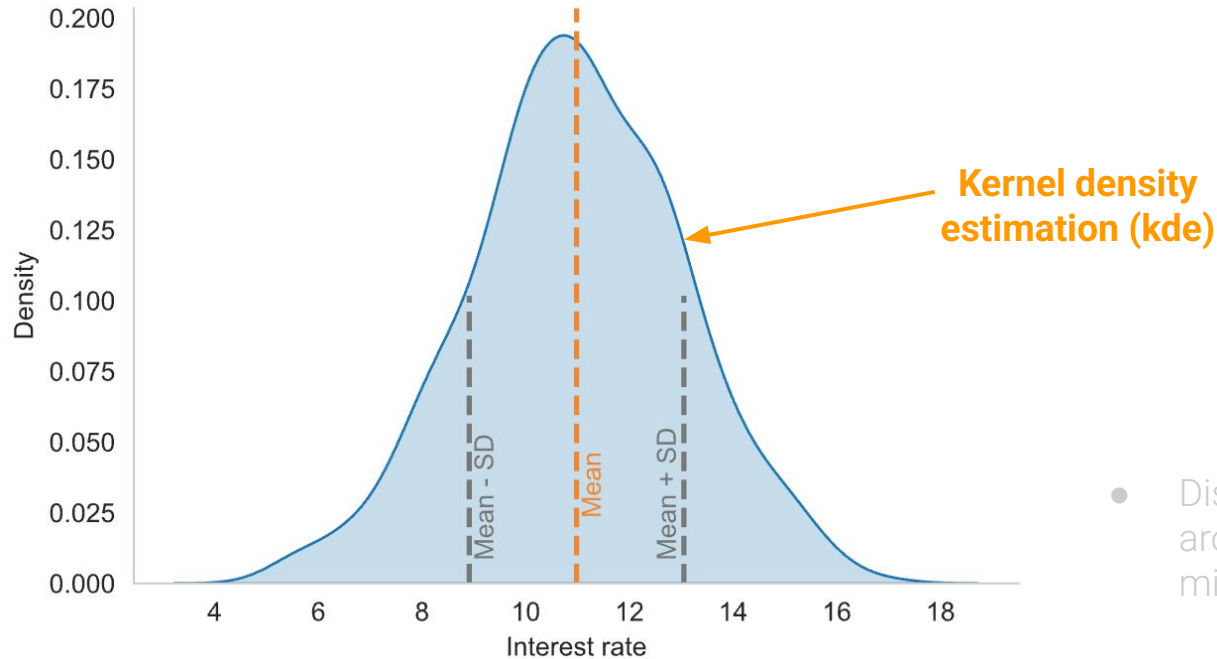
- Distribution is relatively symmetric around a rounded peak in the middle.

Histogram with kernel density estimation (KDE)



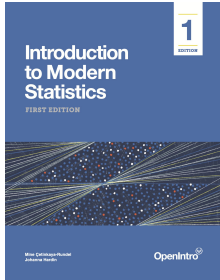
- Distribution is relatively symmetric around a rounded peak in the middle.

Histogram with kernel density estimation (KDE)



- Distribution is relatively symmetric around a rounded peak in the middle.

Resources



The content of this presentation is mainly based on the excellent book “Introduction to Modern Statistics” by Mine Çetinkaya-Rundel and Johanna Hardin (2021).

The online version of the book can be accessed for free:

<https://openintro-ims.netlify.app/index.html>