DATA SCIENCE

# Statistics for Data Science

Central Limit Theorem and Density Estimation



UNC GREENSBORO

# Where We Are

From Data to Probability

So far, we have discussed:

- Random variables
  - Discrete variables
  - Continuous variables
- Probability distributions
- PMF for discrete random variables
- PDF for continuous random variables
- CDF
- Common distributions

Today:

> **We move from knowing a distribution to learning a distribution from data.**

# Data science decisions start with a sample

**Project question**

## Does a new website design increase the average time users spend on the page?

We cannot observe every future user. We randomly assign a sample to Version A or B, measure their sessions, and use the results to guide a decision.

**Population**

All eligible users

**Sample**

Observed A/B participants

**Decision**

Ship, revise, or keep testing

# The sample mean is the project metric

Let $X_i$ be the time on page for user i.

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

**A/B testing compares sample means to estimate a population difference.**

**Version A**

**Average time: 5.2 min**

**Version B**

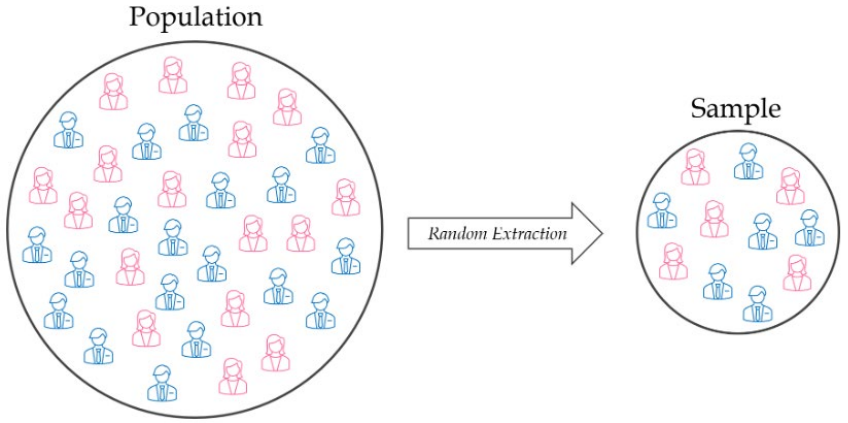
**Average time: 5.6 min**

**Observed difference**

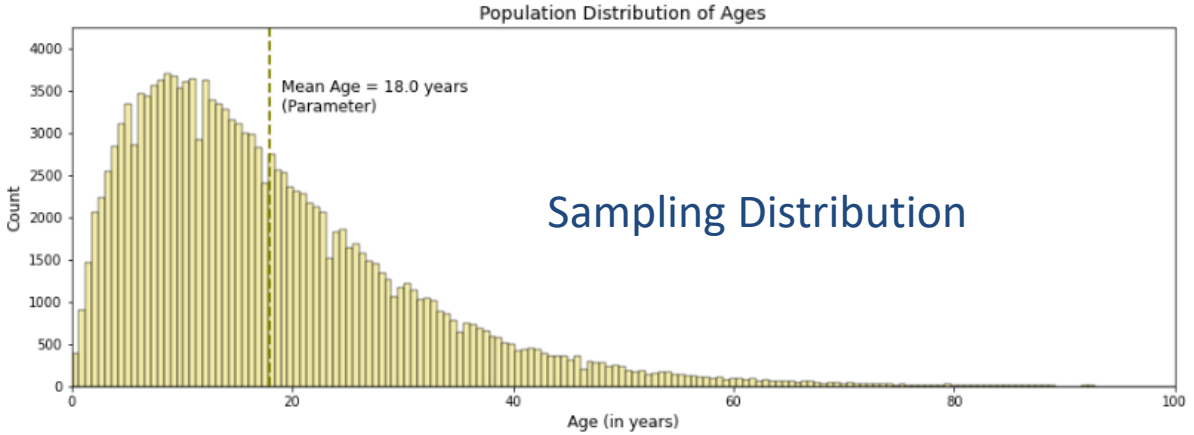
$$\Delta = \bar{X}_A - \bar{X}_B$$

Could random sampling alone explain this difference?

# Central Limit Theorem: Intuition




- Suppose we repeatedly draw random and independent samples from a population.
  - Each sample has size $n$=100
  - For each sample, we calculate the sample mean age
- Example samples:
  - Sample 1 → mean age = 17.8
  - Sample 2 → mean age = 18.4
  - Sample 3 → mean age = 18.1
  - …
  - If we repeat this thousands of times, we obtain thousands of sample means.



# What Does CLT Actually Say?

**The distribution of the sample mean becomes approximately Normal as the sample size increases.**

Population May be:

- Skewed
- Uniform
- Exponential
- Irregular

↓

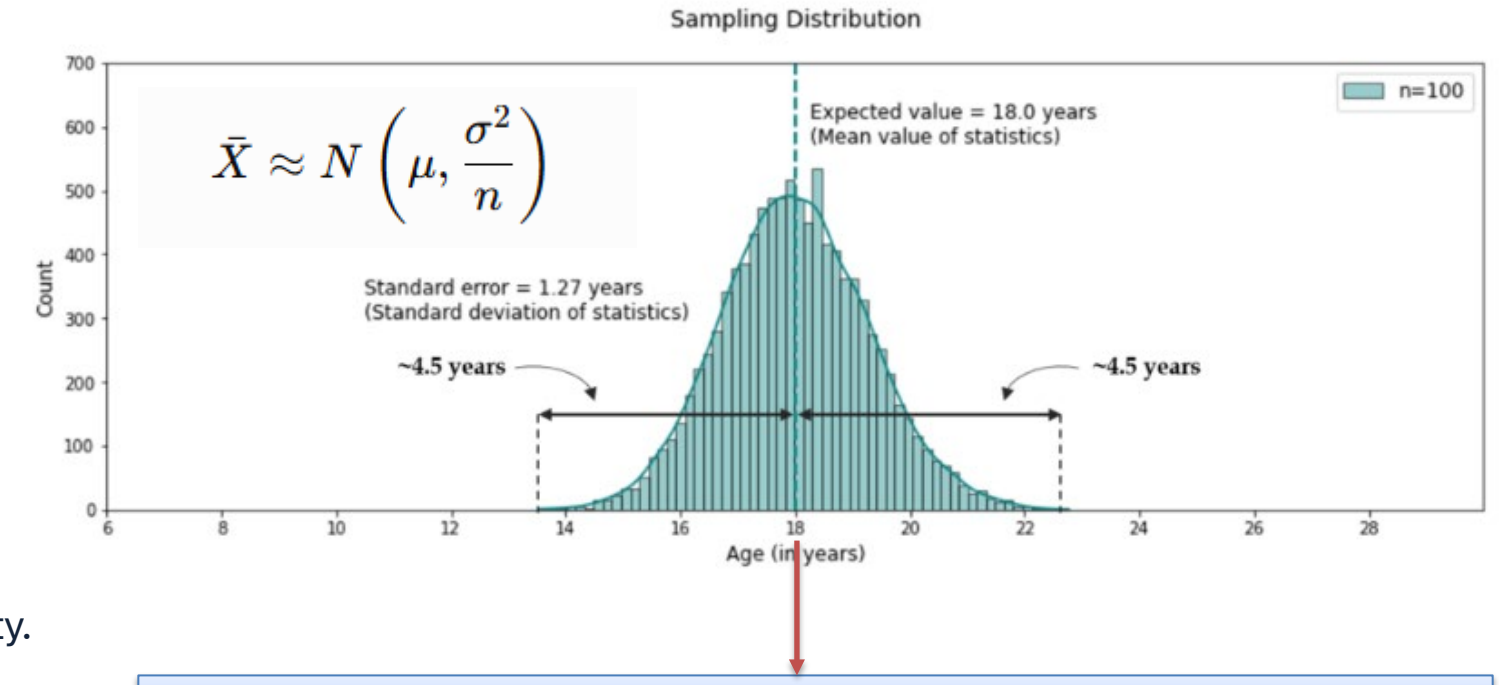
Repeated samples

↓

Calculate one mean from each sample

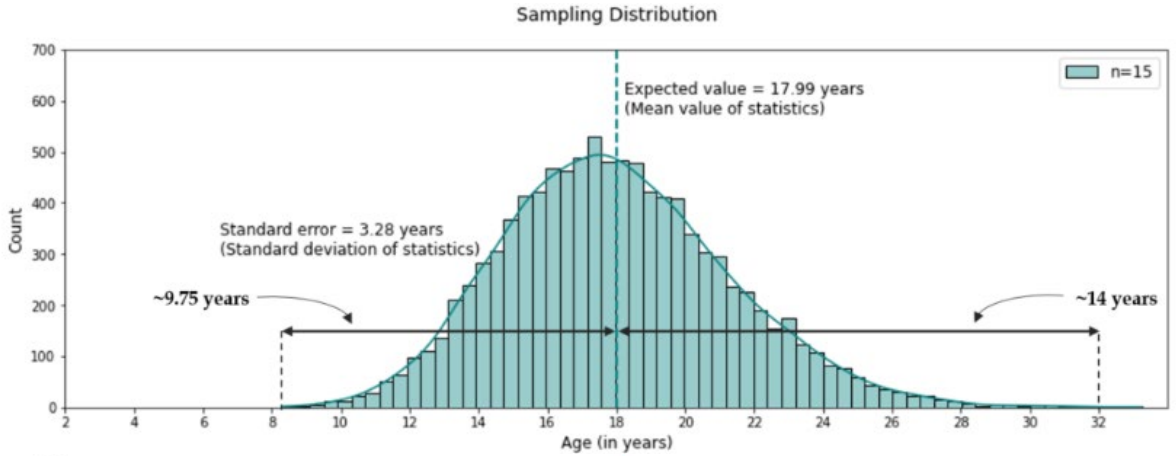
↓

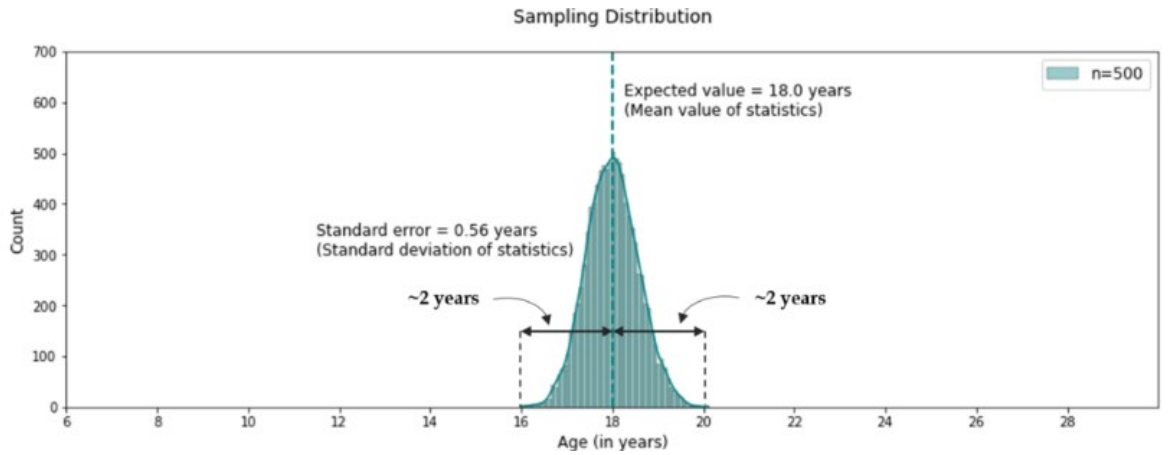
The sampling distribution approaches Normality.

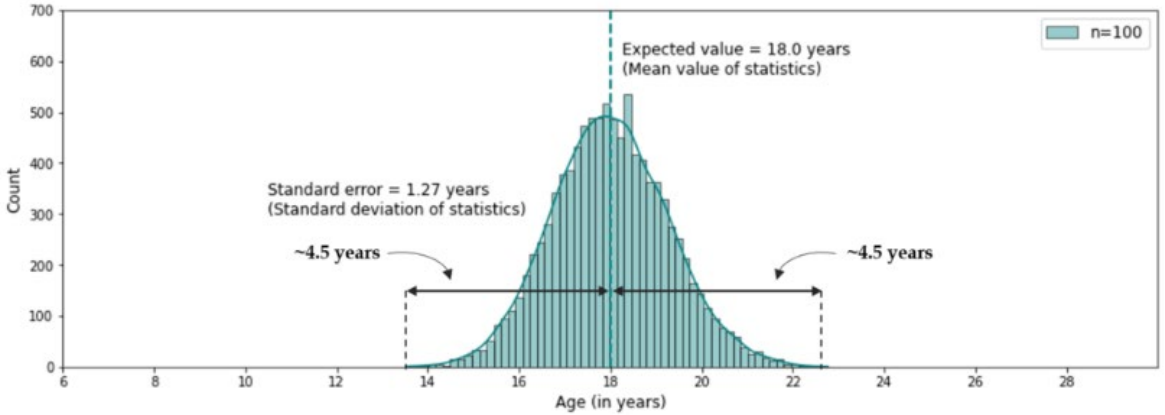


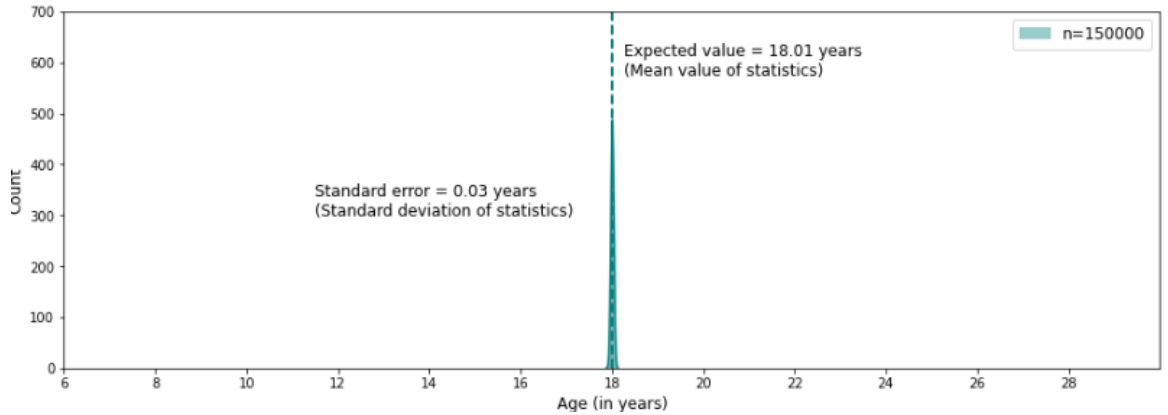

The sample mean is an unbiased estimator of the population mean.

# What Does CLT Actually Say?




As the sample size n increases, the **standard error decreases**, meaning the sample mean becomes more stable.

# The standard error measures uncertainty

**Project question**

## Does a new website design increase the average time users spend on the page?

**Version A**

**Average time: 5.2 min**

**Version B**

**Average time: 5.6 min**

With independent groups, an estimated standard error is

$$SE(\bar{\Delta}) = \sqrt{\frac{{\sigma_A}^2}{N_A} + \frac{{\sigma_B}^2}{N_B}}$$

Suppose each version has 1,000 users. For this worked example, assume $SE(\bar{\Delta})$ = 0.15 min.

# A confidence interval puts a range around the A/B estimate

**Project question**

## Does a new website design increase the average time users spend on the page?

**Version A**

**Average time: 5.2 min**

**Version B**

**Average time: 5.6 min**

Approximate 97% interval

$$\overline{\Delta} \pm 1.96 \times 0.15 \longrightarrow [0.106, 0.694] \text{ minutes}$$

Under the Normal approximation and test assumptions, the interval suggests Version B increases mean time on page.

# A confidence interval puts a range around the A/B estimate

## Project question

**Does a new website design increase th users spend on the page?**

**Version A**

**Average time: 5.2 min**

**Version B**

**Averag**

Entries in the body of the table represents areas under the curve between $-\infty$ and $z$

| $z$ | 0.00 | 0.01 | 0.02 | 0.03 | 0.04 | 0.05 | 0.06 | 0.07 | 0.08 | 0.09 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0.0 | **0.5000** | 0.5040 | 0.5080 | 0.5120 | 0.5160 | 0.5199 | 0.5239 | 0.5279 | 0.5319 | 0.5359 |
| 0.1 | 0.5398 | 0.5438 | 0.5478 | 0.5517 | 0.5557 | 0.5596 | 0.5636 | 0.5675 | 0.5714 | 0.5753 |
| 0.2 | 0.5793 | 0.5832 | 0.5871 | 0.5910 | 0.5948 | 0.5987 | 0.6026 | 0.6064 | 0.6103 | 0.6141 |
| 0.3 | 0.6179 | 0.6217 | 0.6255 | 0.6293 | 0.6331 | 0.6368 | 0.6406 | 0.6443 | 0.6480 | 0.6517 |
| 0.4 | 0.6554 | 0.6591 | 0.6628 | 0.6664 | 0.6700 | 0.6736 | 0.6772 | 0.6808 | 0.6844 | 0.6879 |
| 0.5 | 0.6915 | 0.6950 | 0.6985 | 0.7019 | 0.7054 | 0.7088 | 0.7123 | 0.7157 | 0.7190 | 0.7224 |
| 0.6 | 0.7257 | 0.7291 | 0.7324 | 0.7357 | 0.7389 | 0.7422 | 0.7454 | 0.7486 | 0.7517 | 0.7549 |
| 0.7 | 0.7580 | 0.7611 | 0.7642 | 0.7673 | 0.7704 | 0.7734 | 0.7764 | 0.7794 | 0.7823 | 0.7852 |
| 0.8 | 0.7881 | 0.7910 | 0.7939 | 0.7967 | 0.7995 | 0.8023 | 0.8051 | 0.8078 | 0.8106 | 0.8133 |
| 0.9 | 0.8159 | 0.8186 | 0.8212 | 0.8238 | 0.8264 | 0.8289 | 0.8315 | 0.8340 | 0.8365 | 0.8389 |
| 1.0 | 0.8413 | 0.8438 | 0.8461 | 0.8485 | 0.8508 | 0.8531 | 0.8554 | 0.8577 | 0.8599 | 0.8621 |
| 1.1 | 0.8643 | 0.8665 | 0.8686 | 0.8708 | 0.8729 | 0.8749 | 0.8770 | 0.8790 | 0.8810 | 0.8830 |
| 1.2 | 0.8849 | 0.8869 | 0.8888 | 0.8907 | 0.8925 | 0.8944 | 0.8962 | 0.8980 | 0.8997 | 0.9015 |
| 1.3 | 0.9032 | 0.9049 | 0.9066 | 0.9082 | 0.9099 | 0.9115 | 0.9131 | 0.9147 | 0.9162 | 0.9177 |
| 1.4 | 0.9192 | 0.9207 | 0.9222 | 0.9236 | 0.9251 | 0.9265 | 0.9279 | 0.9292 | 0.9306 | 0.9319 |
| 1.5 | 0.9332 | 0.9345 | 0.9357 | 0.9370 | 0.9382 | 0.9394 | 0.9406 | 0.9418 | 0.9429 | 0.9441 |
| 1.6 | 0.9452 | 0.9463 | 0.9474 | 0.9484 | **0.9495** | **0.9505** | 0.9515 | 0.9525 | 0.9535 | 0.9545 |
| 1.7 | 0.9554 | 0.9564 | 0.9573 | 0.9582 | 0.9591 | 0.9599 | 0.9608 | 0.9616 | 0.9625 | 0.9633 |
| 1.8 | 0.9641 | 0.9649 | 0.9656 | 0.9664 | 0.9671 | 0.9678 | 0.9686 | 0.9693 | 0.9699 | 0.9706 |
| 1.9 | 0.9713 | 0.9719 | 0.9726 | 0.9732 | 0.9738 | 0.9744 | **0.9750** | 0.9756 | 0.9761 | 0.9767 |

Approximate 97% interval

$$\overline{\Delta} \pm 1.96 \times 0.15 \longrightarrow [0.106, 0.694] \text{ minutes}$$

Under the Normal approximation and test assumptions, the interval suggests Version B increases mean time on page.

# In a project, the distribution is usually unknown

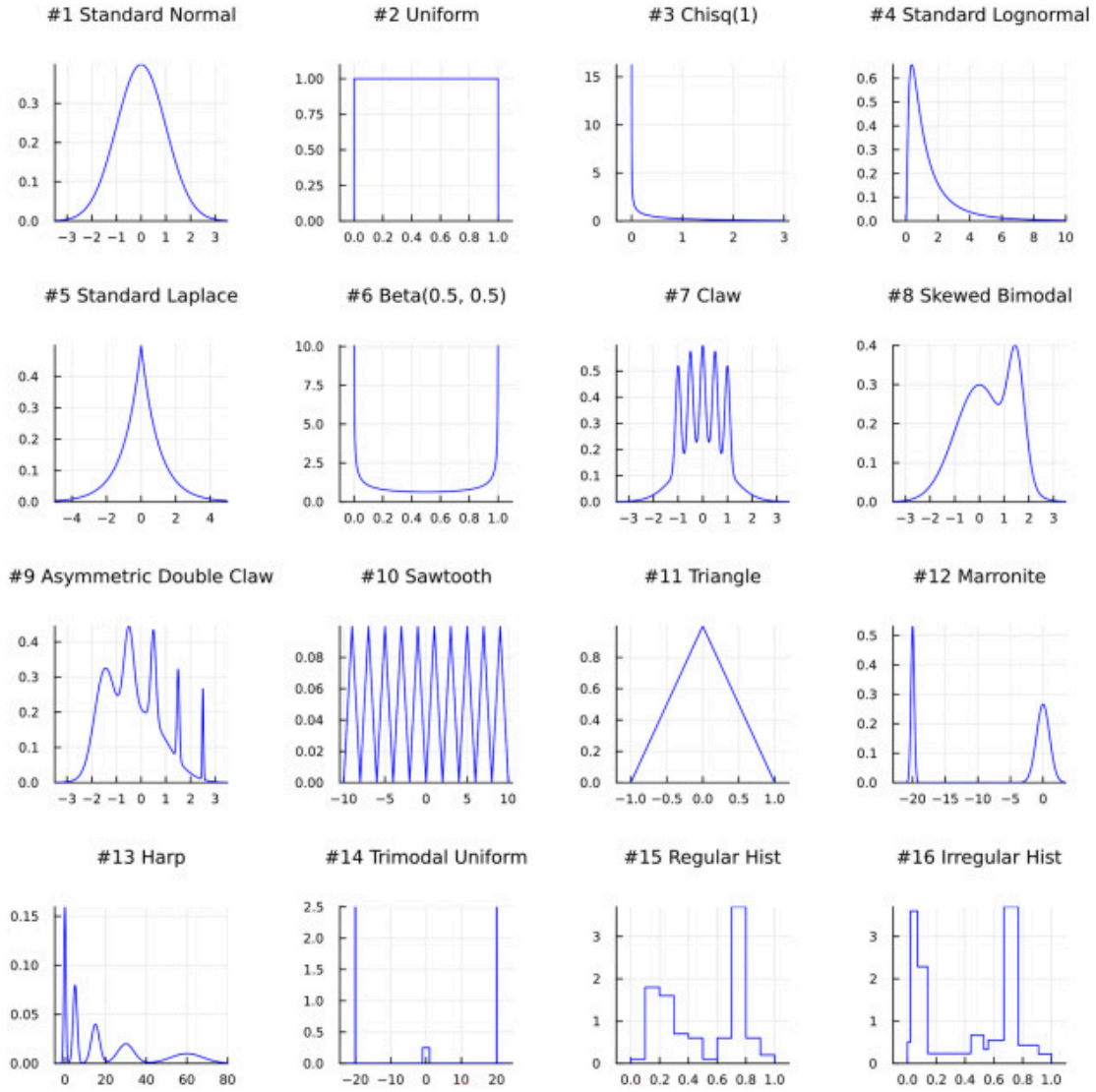
Suppose we observe:

```
Response time (ms)

20  22  25  21  23  28  30
...
150 200 450
```

If we know the distribution, we can ask:

- What is the probability that response time > 100 ms?
- What is the 95th percentile?
- What counts as an unusual response?
- What threshold should trigger an alert?



# Why Is Knowing the Distribution Important?

A probability distribution summarizes how likely different values are.

Knowing the distribution helps us:

- Understand the data
- Calculate probabilities
- Identify unusual observations
- Estimate percentiles
- Simulate new observations
- Build statistical models
- Detect anomalies
- Make predictions
- Quantify uncertainty

# Use observed data to estimate the underlying distribution

**Previously**

**If $X \sim \text{Normal}(\mu, \sigma^2)$, we can calculate probabilities from its PDF.**

**In practice**

**We often have observations $x_1, \ldots, x_n$ but do not know the population distribution.**

## Goal: use observed data to estimate the underlying density f(x).

This is the density estimation problem.

# Parametric vs. Non-Parametric

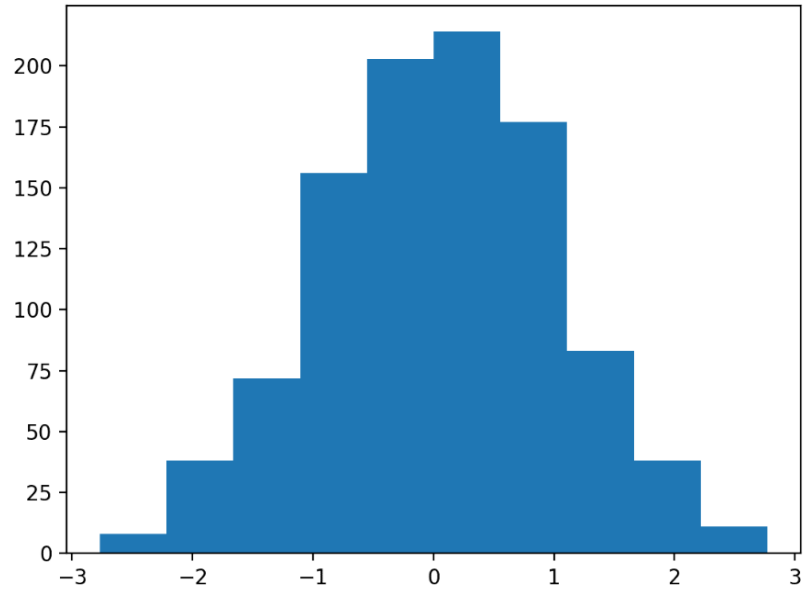
There are two major approaches:

- **Parametric Density Estimation**: Assume a particular family of distributions.
- **Non-Parametric Density Estimation**: Make fewer assumptions about the distribution.

| | Parametric | Non-parametric |
|---|---|---|
| **Assumption** | Choose a family, such as Normal | Make fewer shape assumptions |
| **What is estimated** | A fixed set of parameters | A density that can adapt to the data |
| **Example** | Normal model | Histogram or KDE |
| **Strength** | Compact and interpretable | Can capture skew or multiple peaks |
| **Risk** | Wrong family misses structure | Needs data and tuning |

# Parametric Density Estimation

Suppose the spending data look roughly Normal.




$$X \sim \text{Normal}(\mu, \sigma^2)$$

**The model is described by $\mu$ and $\sigma^2$.**

Estimate those parameters from the sample, then use the fitted distribution for probabilities or percentiles.

**Choose a family → estimate parameters → fitted density**

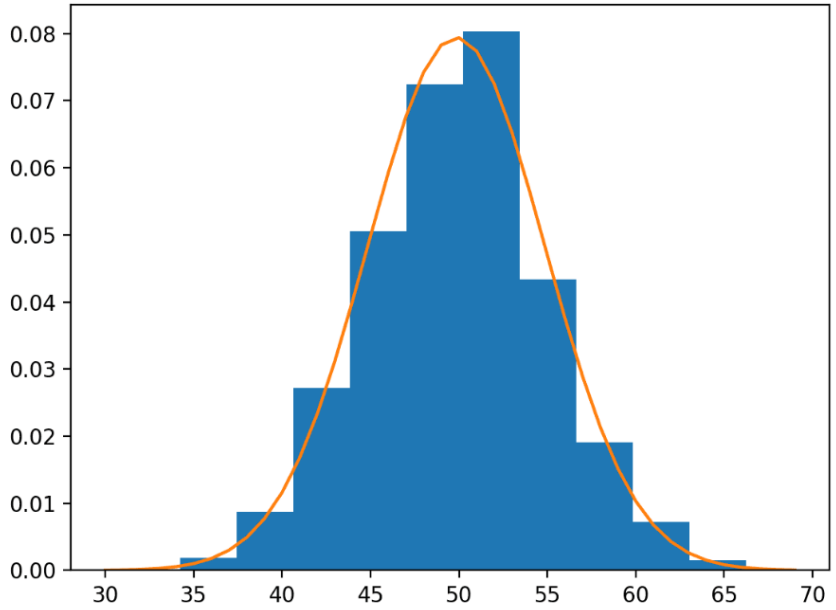
# Example: Modeling Exam Scores

Suppose we observe:

```
72, 75, 81, 68, 90,
77, 84, 73, 79, 88, ...
```

We might hypothesize:

$$X \sim N(\mu, \sigma^2)$$




Normal Distribution: parameters can be estimated from data by calculating the **sample mean** and **sample standard deviation**.

**Check if it is a good fit:**

- Plotting the density function and comparing the shape to the histogram.
- Sampling the density function and comparing the generated sample to the real sample.
- Using a statistical test to confirm the data fits the distribution.

# Method of Moments (MoM)

The Method of Moments estimates unknown parameters by matching:

**Theoretical moments = Sample moments**

Normal Distribution

$$X \sim \mathcal{N}(\mu, \sigma^2)$$

Theoretical Moments

Mean: $\mathbb{E}[X] = \mu$ ← $\bar{x} = \frac{1}{n}\sum x_i$

Variance: $\mathrm{Var}(X) = \sigma^2$ ← $\frac{1}{n}\sum (x_i - \bar{x})^2$

Exponential Distribution

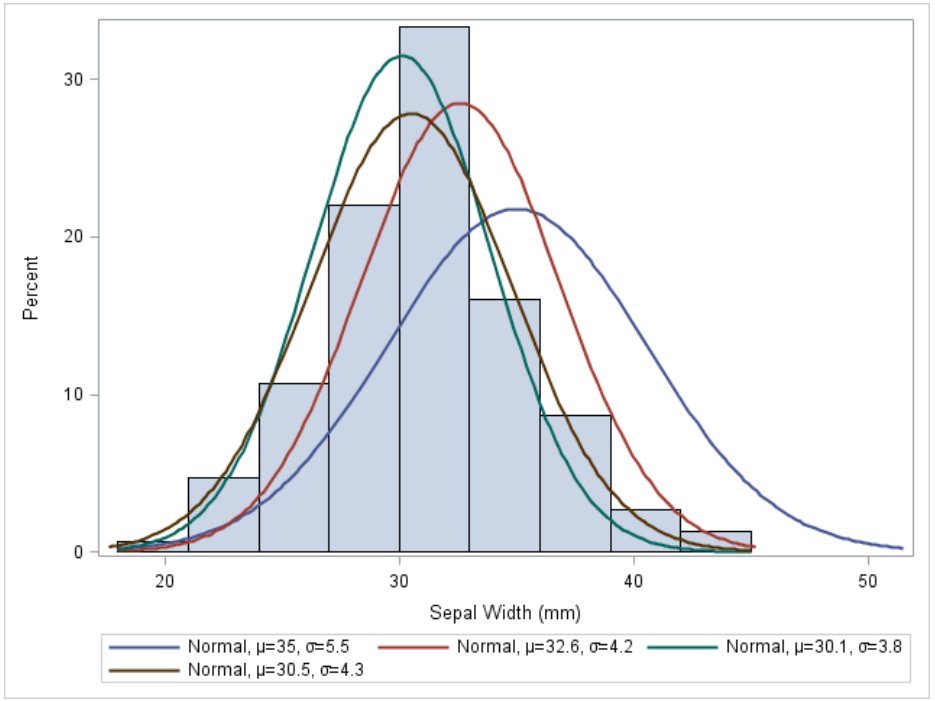
$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0, \\ 0 & x < 0. \end{cases}$$

Theoretical Moments

Mean: $\mathbb{E}[X] = \frac{1}{\lambda}$ ← $\bar{x} = \frac{1}{n}\sum x_i$

$$\hat{\lambda} = \frac{1}{\bar{x}}$$

# Strengths and Limitations of MoM




### ADVANTAGES

- Simple
- Intuitive
- Often computationally easy
- Useful when moments have simple formulas

### LIMITATIONS

- Estimates may not be optimal
- Moments may not exist
- Different parameterizations can produce complicated equations
- Does not directly ask which parameter values make the observed data most likely

**We need a general method to estimate parameters from data. This motivates another important method:**

> **Maximum Likelihood Estimation**

# Maximum Likelihood Estimation (MLE)

Maximum Likelihood Estimation (MLE): Adjust parameters to maximize the probability of observed data.

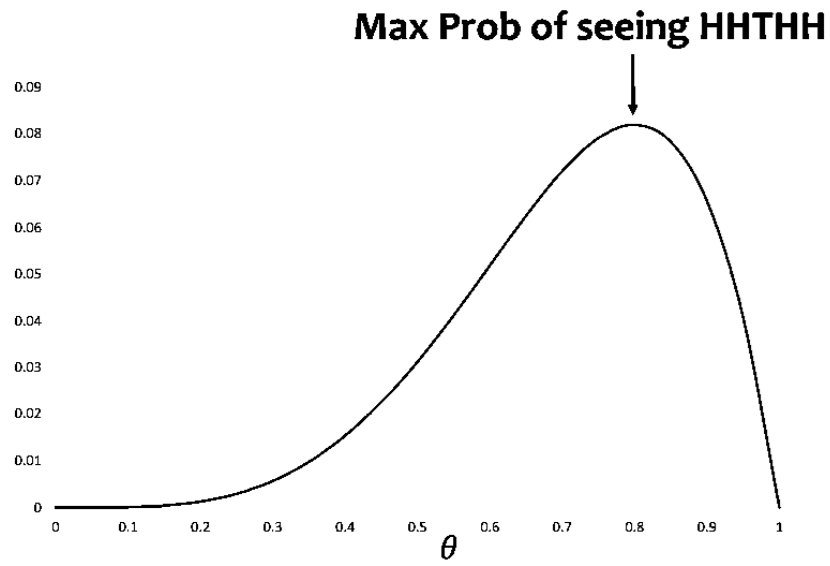
Suppose we have a mystery coin with some probability $\theta$ of coming up heads. We flip the coin 5 times, independent of other flips.

## Likelihood

Say we see outcome HHTHH.

$$\mathcal{L}(HHTHH \mid \theta) = \theta^4(1-\theta)$$

Probability of observing the outcome HHTHH if $\theta$ = prob. of heads. This is a function of $\theta$.



# Maximum Likelihood Estimation (MLE)

Maximum Likelihood Estimation (MLE): Adjust parameters to maximize the probability of observed data.

**Definition.** The **likelihood** of independent observations $x_1, \ldots, x_n$ is

$$\mathcal{L}(x_1, \ldots, x_n|\theta) = \prod_{i=1}^{n} \mathbb{P}(x_i; \theta)$$

**Maximum Likelihood Estimation (MLE).** Given data $x_1, \ldots, x_n$, find $\hat{\theta}$ ("the MLE") of model such that $L(x_1, \ldots, x_n|\hat{\theta})$ is maximized!

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, \mathcal{L}(x_1, \ldots, x_n|\theta)$$

Solve $\frac{\partial L(x_1, \ldots, x_n|\theta)}{\partial \theta} = 0$

# Maximum Likelihood Estimation (MLE)

Maximum Likelihood Estimation (MLE): Adjust parameters to maximize the probability of observed data.

Observe: Coin-flip outcomes $x_1, \ldots, x_n$, with $n_H$ heads, $n_T$ tails

– I.e., $n_H + n_T = n$

**Goal:** estimate $\theta$ = prob. heads.

$$L(x_1, \ldots, x_n|\theta) = \theta^{n_H}(1-\theta)^{n_T}$$

$$\frac{\partial}{\partial\theta} L(x_1, \ldots, x_n|\theta) = ???$$

While it is possible to compute this derivative, it's not always nice since we are working with products.

# Maximum Likelihood Estimation (MLE)

Maximum Likelihood Estimation (MLE): Adjust parameters to maximize the probability of observed data.

**Definition.** The **log-likelihood** of independent observations $x_1, \ldots, x_n$ is

$$\mathcal{LL}(x_1, \ldots, x_n|\theta) = \ln \mathcal{L}(x_1, \ldots, x_n|\theta)$$

$$= \ln \prod_{i=1}^{n} \mathbb{P}(x_i; \theta) = \sum_{i=1}^{n} \ln \mathbb{P}(x_i; \theta)$$

# Maximum Likelihood Estimation (MLE)

Maximum Likelihood Estimation (MLE): Adjust parameters to maximize the probability of observed data.

Observe: Coin-flip outcomes $x_1, \ldots, x_n$, with $n_H$ heads, $n_T$ tails

– I.e., $n_H + n_T = n$

**Goal:** estimate $\theta$ = prob. heads.

$$\mathcal{L}(x_1, \ldots, x_n|\theta) = \theta^{n_H}(1-\theta)^{n_T}$$

$$\ln \mathcal{L}(x_1, \ldots, x_n|\theta) = n_H \ln \theta + n_T \ln(1-\theta)$$

$$\frac{\partial}{\partial \theta} \ln \mathcal{L}(x_1, \ldots, x_n|\theta) = n_H \cdot \frac{1}{\theta} - n_T \cdot \frac{1}{1-\theta}$$

Solve $n_H \cdot \frac{1}{\hat{\theta}} - n_T \cdot \frac{1}{1-\hat{\theta}} = 0$

$$\hat{\theta} = \frac{n_H}{n}$$

**Useful log properties:**

$$\log(ab) = \log(a) + \log(b)$$
$$\log(a/b) = \log(a) - \log(b)$$
$$\log(a^b) = b\log(a)$$

# Maximum Likelihood Estimation (MLE)

Test your understanding…

I have a bag that contains $3$ balls. Each ball is either red or blue, but I have no information in addition to this. Thus, the number of blue balls, call it $\theta$, might be $0$, $1$, $2$, or $3$. I am allowed to choose $4$ balls at random from the bag with replacement. We define the random variables $X_1$, $X_2$, $X_3$, and $X_4$ as follows

$$X_i = \begin{cases} 1 & \text{if the } i\text{th chosen ball is blue} \\ 0 & \text{if the } i\text{th chosen ball is red} \end{cases}$$

Note that $X_i$'s are i.i.d. and $X_i \sim Bernoulli(\frac{\theta}{3})$. After doing my experiment, I observe the following values for $X_i$'s.

$$x_1 = 1, x_2 = 0, x_3 = 1, x_4 = 1.$$

Thus, I observe $3$ blue balls and $1$ red balls.

1. For each possible value of $\theta$, find the probability of the observed sample, $(x_1, x_2, x_3, x_4) = (1, 0, 1, 1)$.
2. For which value of $\theta$ is the probability of the observed sample is the largest?

# MoM vs. MLE

| | Method of Moments | Maximum Likelihood |
|---|---|---|
| **Question** | Do model moments match the sample? | Which parameters make the data most likely? |
| **Method** | Set theoretical moments equal to sample moments | Maximize likelihood or log-likelihood |
| **Computation** | Often direct | May need optimization |
| **Data science link** | Useful baseline method | Common in probabilistic models |

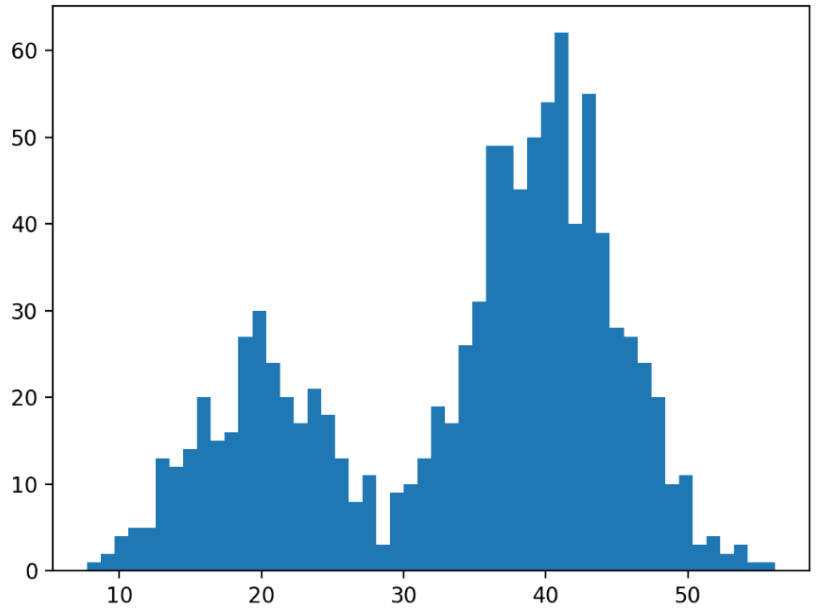
Important:

For some models, MoM and MLE give the same answer.

For others, they can produce different estimates.

# Non-Parametric Density Estimation

In some cases, a data sample may not resemble a common probability distribution or cannot be easily made to fit the distribution.




This is often the case when the data has two peaks (bimodal distribution) or many peaks (multimodal distribution)

# Histogram as a Simple Density Estimate

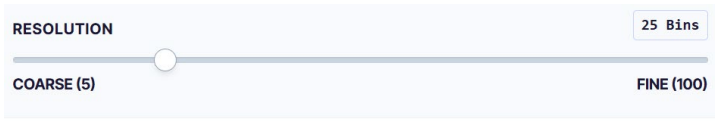
The Binning Trap: Histograms seem simple, but they can be misleading.

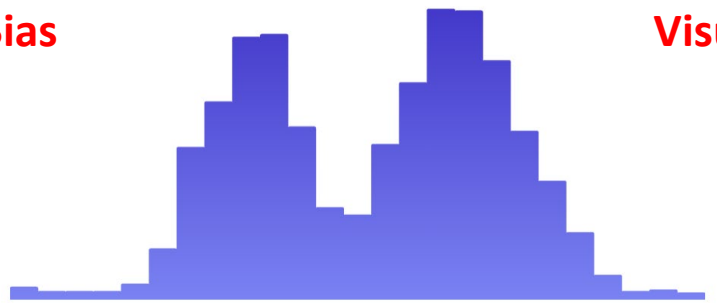





looks unimodal

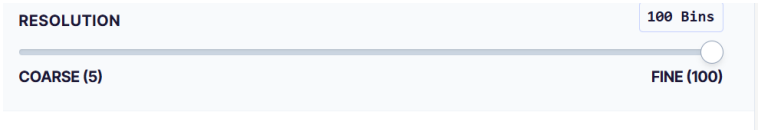
Visual Bias



looks bimodal

Visual Bias






looks noisy

**Bin Width:**
Too small → noisy, overfitting
Too large → oversmoothed, lose structure

So how can we avoid these arbitrary bins?

https://www.nveil.com/histogram-history-binning-bias-kde/

# Kernel Density Estimation (KDE)

KDE estimates the density by placing a small smooth curve around every observation.

Instead of:

> **"How many points fall inside this bin?"**

we ask:

> **"How much does each observation contribute to the density around it?"**

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

where,

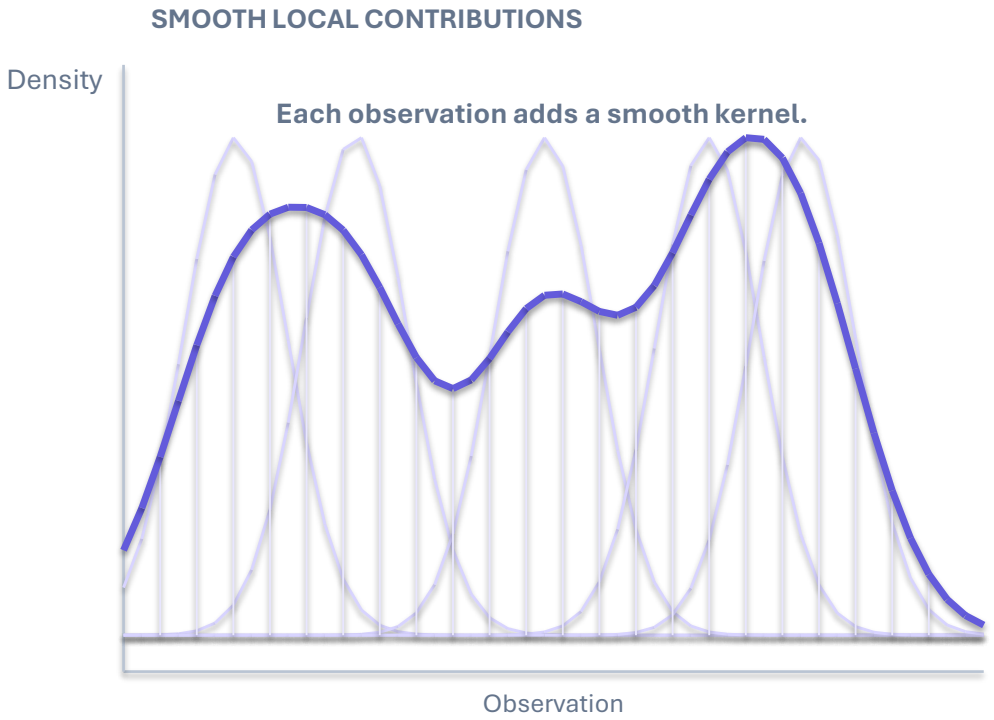
**K ->** kernel (non-negative function)

**h ->** bandwidth (smoothing parameter, h > 0)

$K_{h\ ->}$ scaled kernel

$f_{h(x)\ ->}$ density (to calculate)

**n ->** no. of samples in random sample.



# Kernel Density Estimation (KDE)

Imagine each observation is a small "hill."

For example:

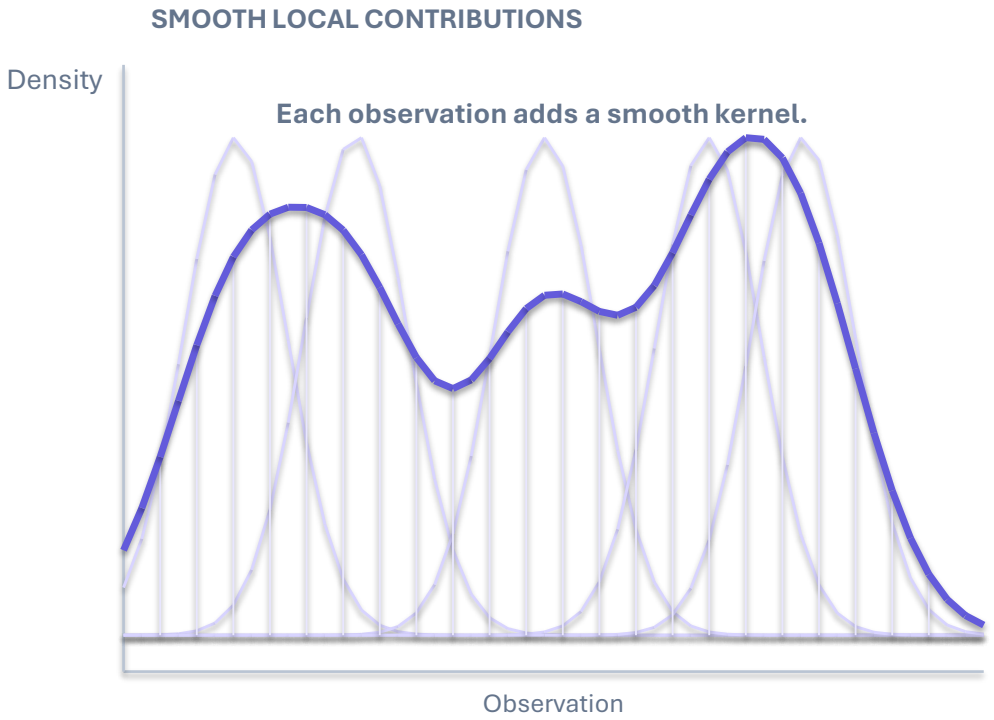
```
Observation: 10

          /\
        /    \
      /        \
-----/----------\-----
      8   10   12
```

For many observations:

```
     /\        /\
    /  \  /\  /  \
___/____\/__\/____\____
```

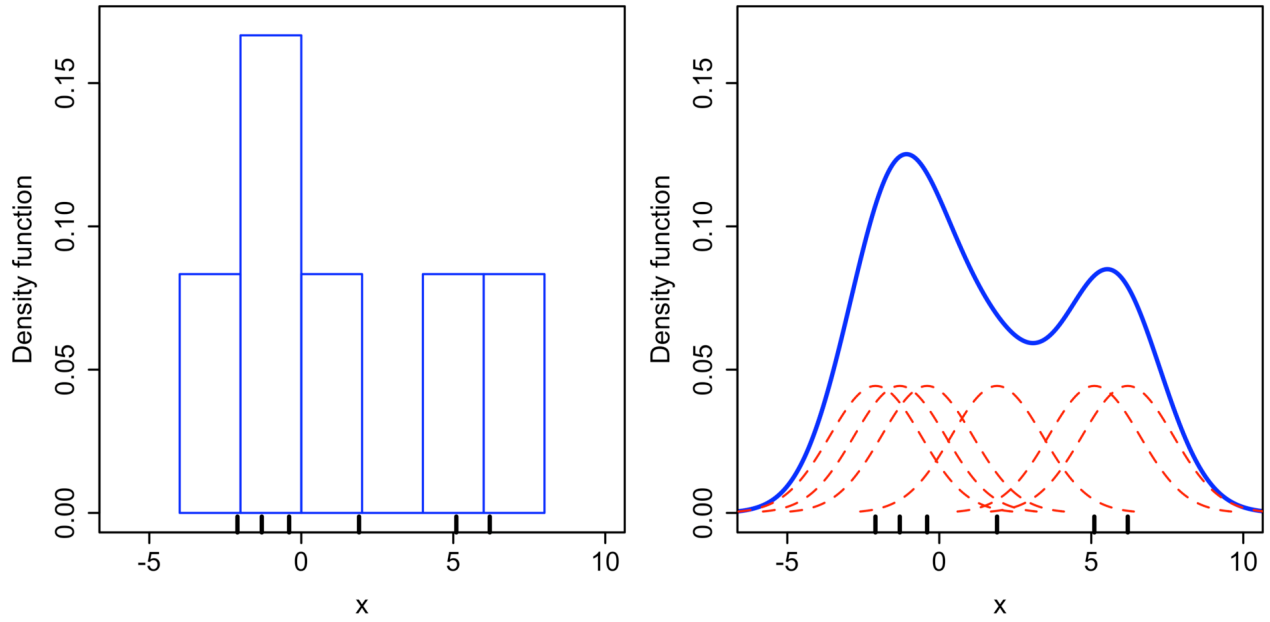
Add all the small hills together.
The result is the estimated density.




**KDE = many local contributions → one smooth density estimate**

# Kernel Density Estimation (KDE)

| Sample | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| Value | -2.1 | -1.3 | -0.4 | 1.9 | 5.1 | 6.2 |



# Kernel Density Estimation (KDE)

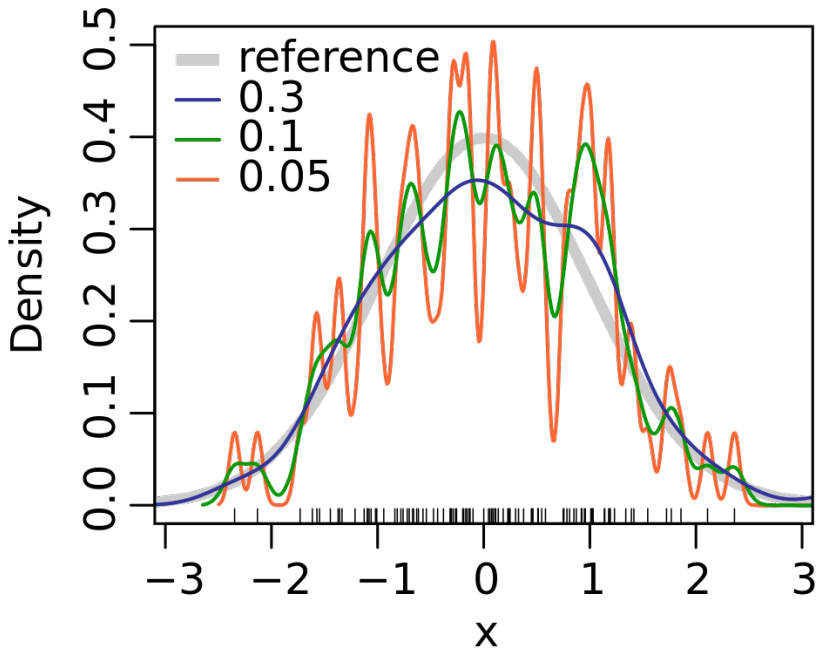
A kernel determines the shape of the contribution from each observation.

A common choice is the Gaussian kernel:

$$K(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2}$$

Each observation contributes most strongly near itself.

The exact kernel shape is often less important than the choice of bandwidth.



# KDE: Under-Smoothing vs. Over-Smoothing

Too small bandwidth



May capture noise.

Good bandwidth



Captures the general structure.

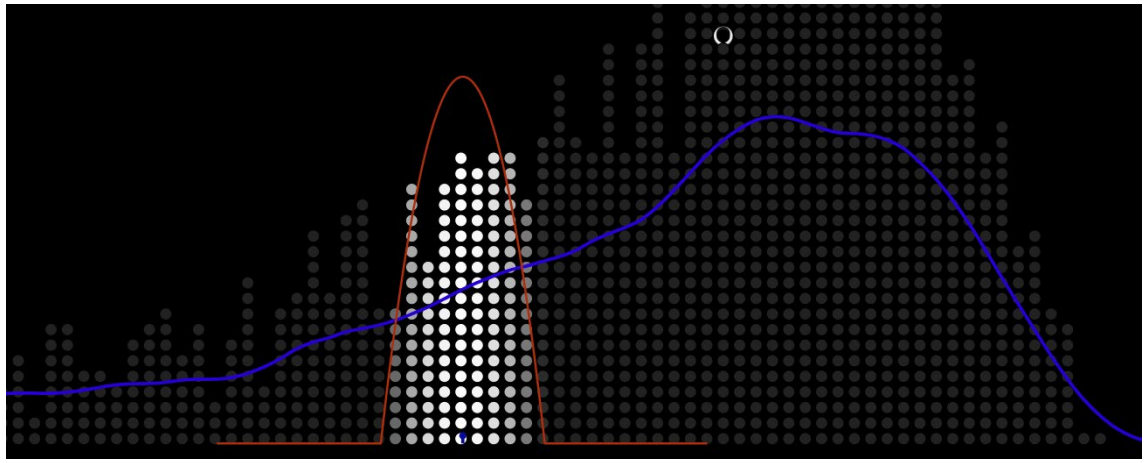
Too large bandwidth



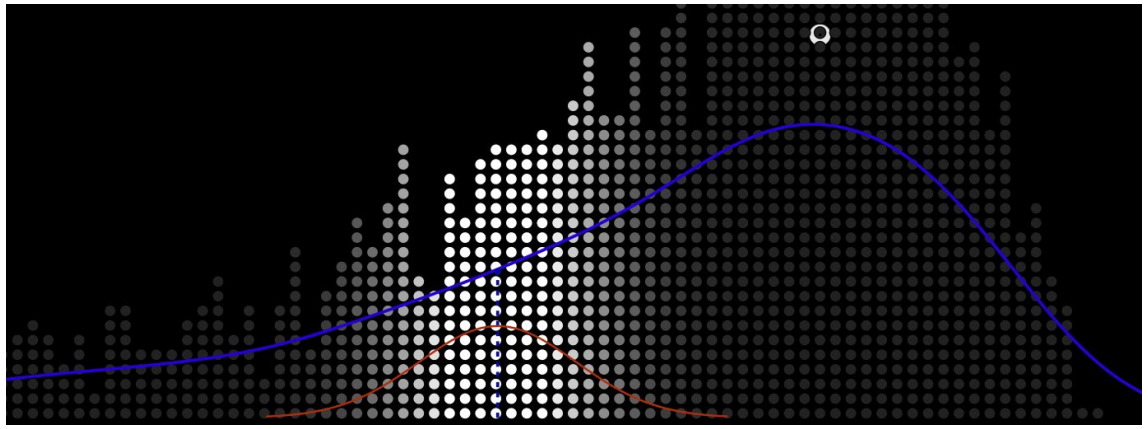
May hide important features.

# Kernel Density Estimation (KDE)

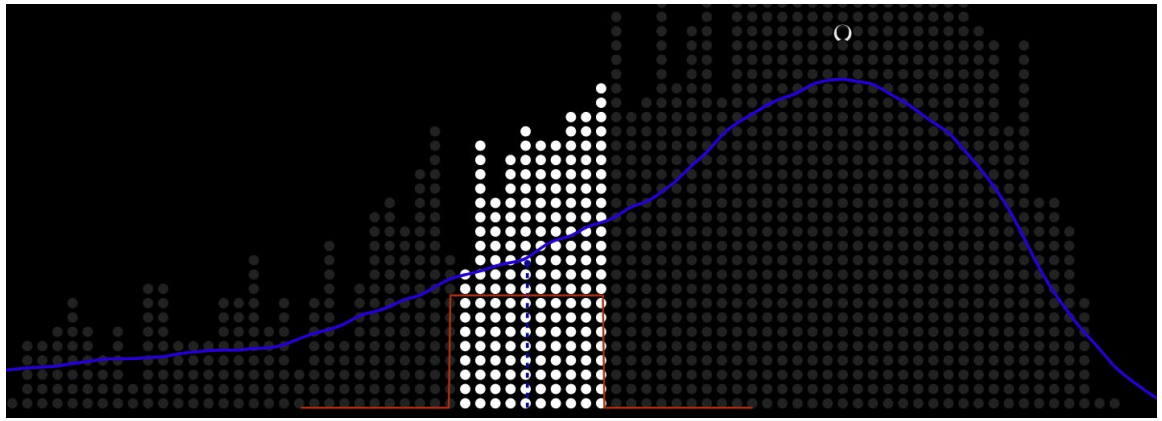
Kernel: Epanechnikov
Bandwidth: 0.05
Amplitude: 3.00



Kernel: Normal
Bandwidth: 0.05
Amplitude: 3.00



Kernel: Uniform
Bandwidth: 0.05
Amplitude: 3.00



https://mathisonian.github.io/kde/

# Density Estimation in Python

Suppose we have:

```
import pandas as pd
df = pd.read_csv("customers.csv")
spending = df["monthly_spending"]
```

Using pandas:

```
spending.plot(kind="density")
```

Or using seaborn:

```
import seaborn as sns
sns.kdeplot(data=spending)
```