Stat C131A: Statistical Methods for Data Science

Lecture 7: Probability Distributions

a) Kernel Density Estimation, review
b) The Central Limit Theorem

Sep 18 2025

Kernel Density Estimation: KDE

- In kernel density estimation, the estimate is $\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i)$
    - K is the Kernel function

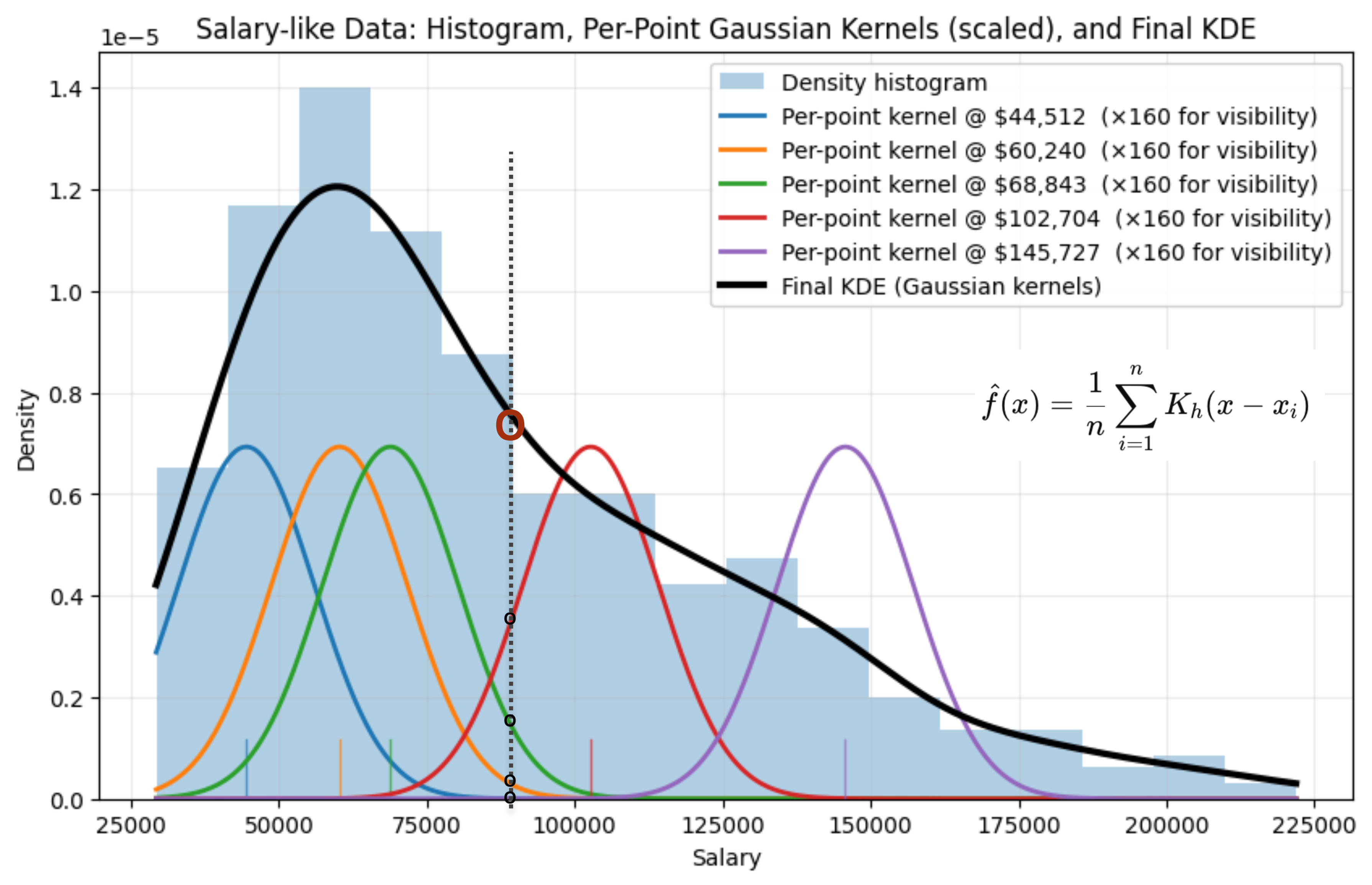
- For the Gaussian kernel, K is the normal distribution PDF.

- **Each data point x_i contributes a normal curve** with
    - **mean = x_i** and
    - **standard deviation = h** (the bandwidth).

- The bandwidth h controls the spread of each normal curve
    - small h gives narrow bumps (wiggly estimate), large h gives wide bumps (smoother estimate).

- The **final KDE is the average of all these normal curves** across the data points.

KDE



Salary-like Data: Histogram, Per-Point Gaussian Kernels (scaled), and Final KDE

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i)$$

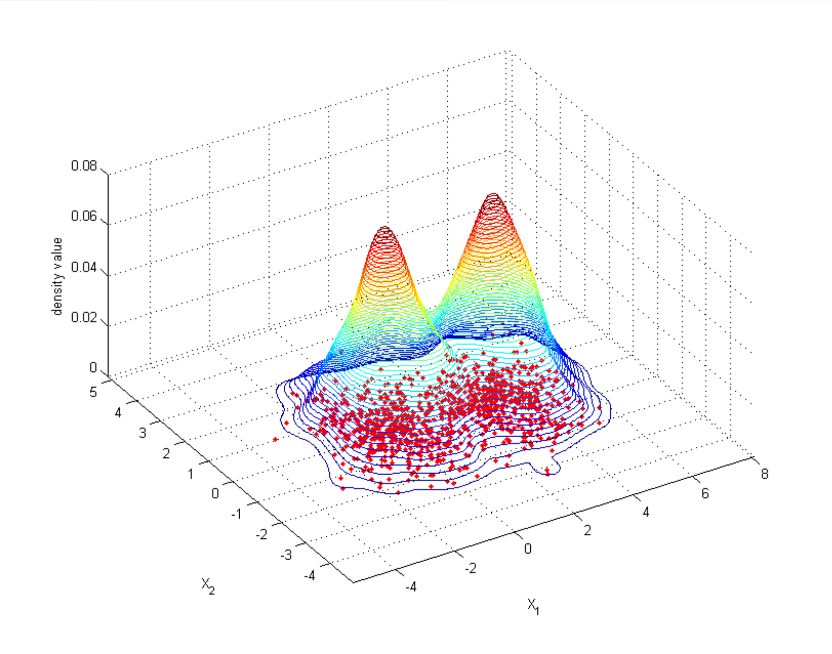
Choice of Kernel

- Common kernels: Gaussian, Epanechnikov, Uniform
- All valid as long as integrate to 1
- Shape less important than bandwidth !
  - Why ?

Bandwidth (h)

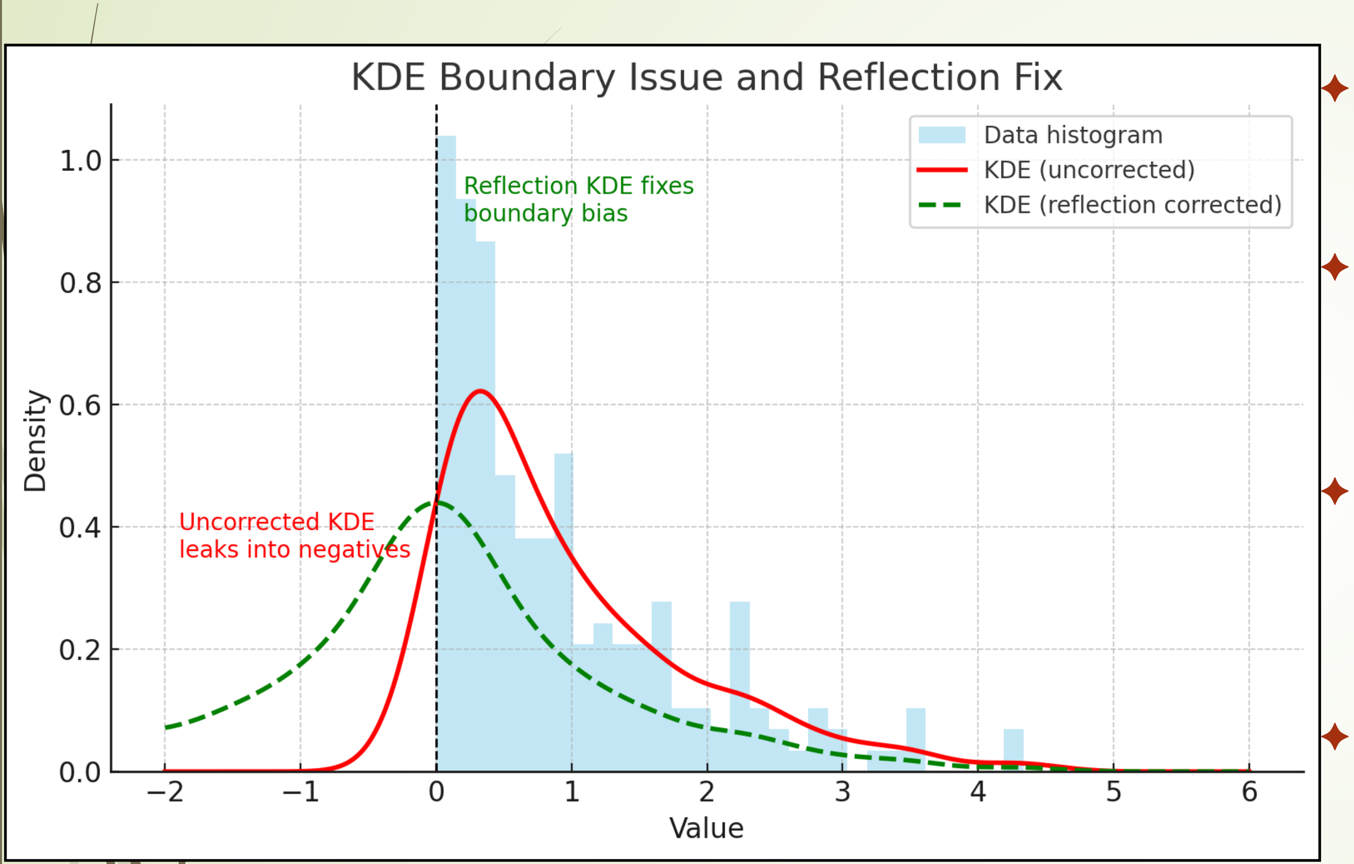
- Controls **smoothness** of KDE
- Small h: **very wiggly** (overfit)
- Large h: **very smooth** (underfit)

Multivariate KDE



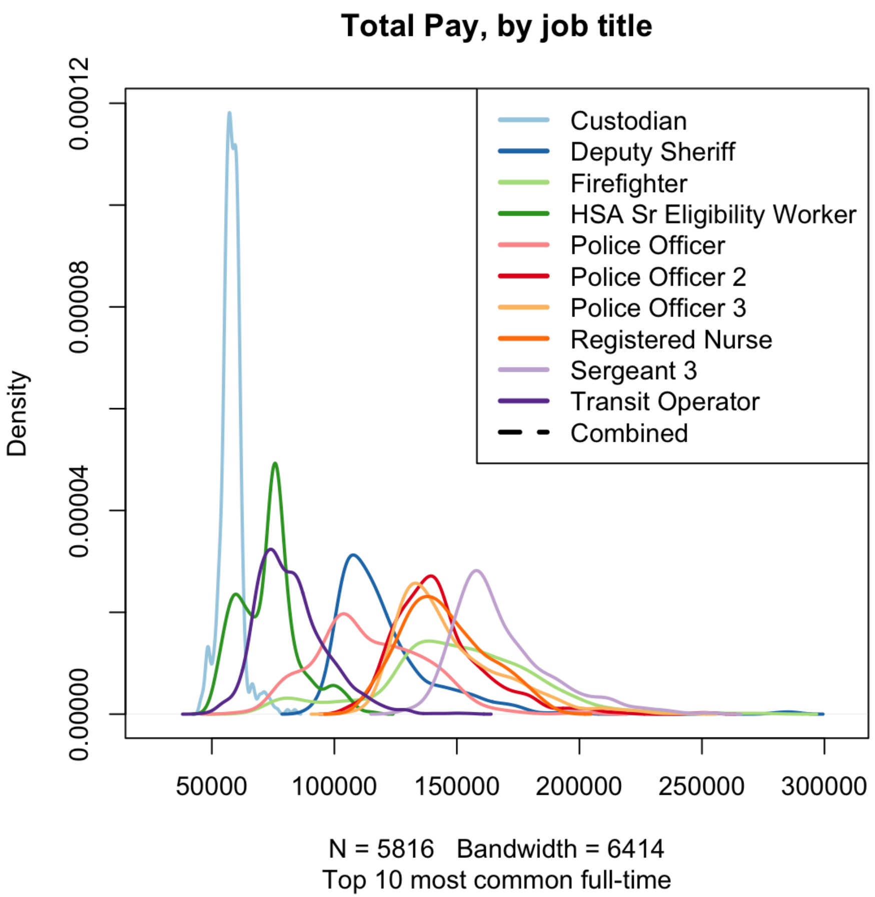
- Extension to higher dimensions
- Use **product of kernels**
- Bandwidth **matrix** controls smoothness in each direction

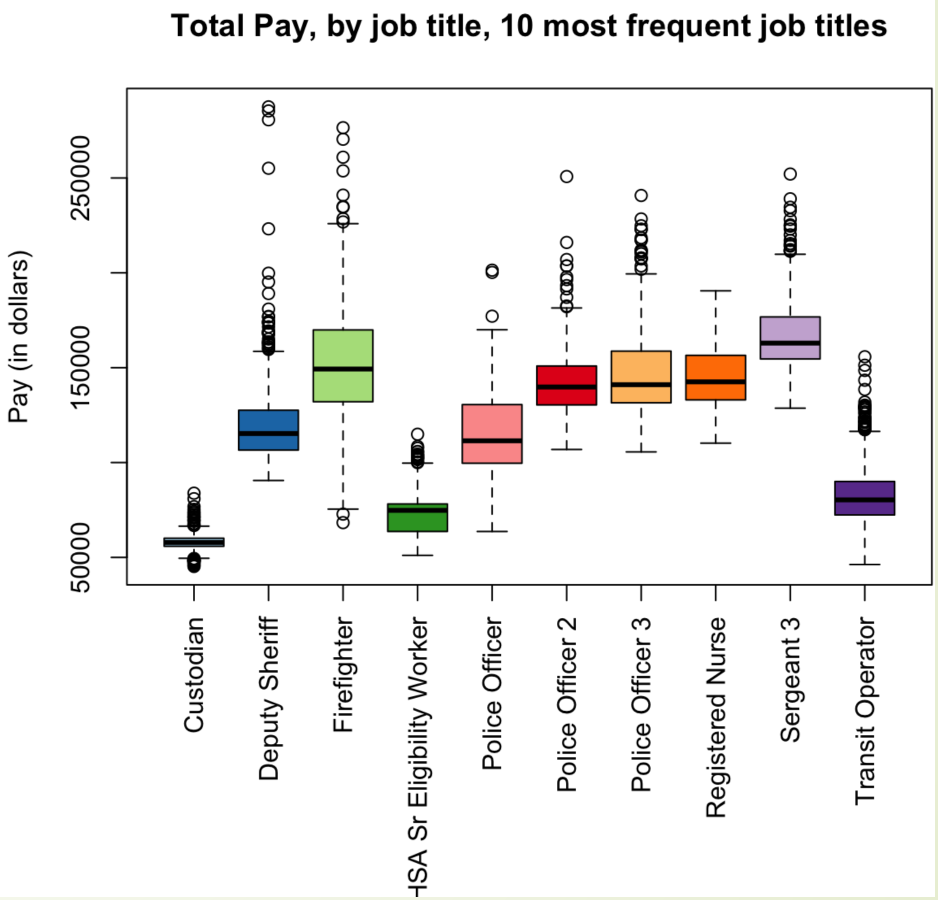
Boundary Issues



- When data are naturally bounded (e.g., incomes ≥ 0, probabilities ≤ 1), kernels **extend beyond those boundaries**.
- At the edges, Gaussian kernels (and others) place weight outside the valid range, “leaking” density into impossible regions.
- This causes the estimated PDF near boundaries to be biased downward.
  - **Underestimates density at the edges**.
- The effect is most visible when many data points lie near the boundary (e.g., 0 values) !
- Solutions: **boundary-corrected kernels**, **reflection methods** (mirror data at boundaries), or transforming the data to an unbounded scale.

Comparing Groups with Density Curves





The **Central Limit Theorem**

Sample Statistics

- The probability distribution of **all the possible values a sample statistic can take** is called the sampling distribution of the statistic.
  - The key word here is **"sample statistic"**

- Sample mean and sample proportion based on a random sample are examples of sample statistic(s).

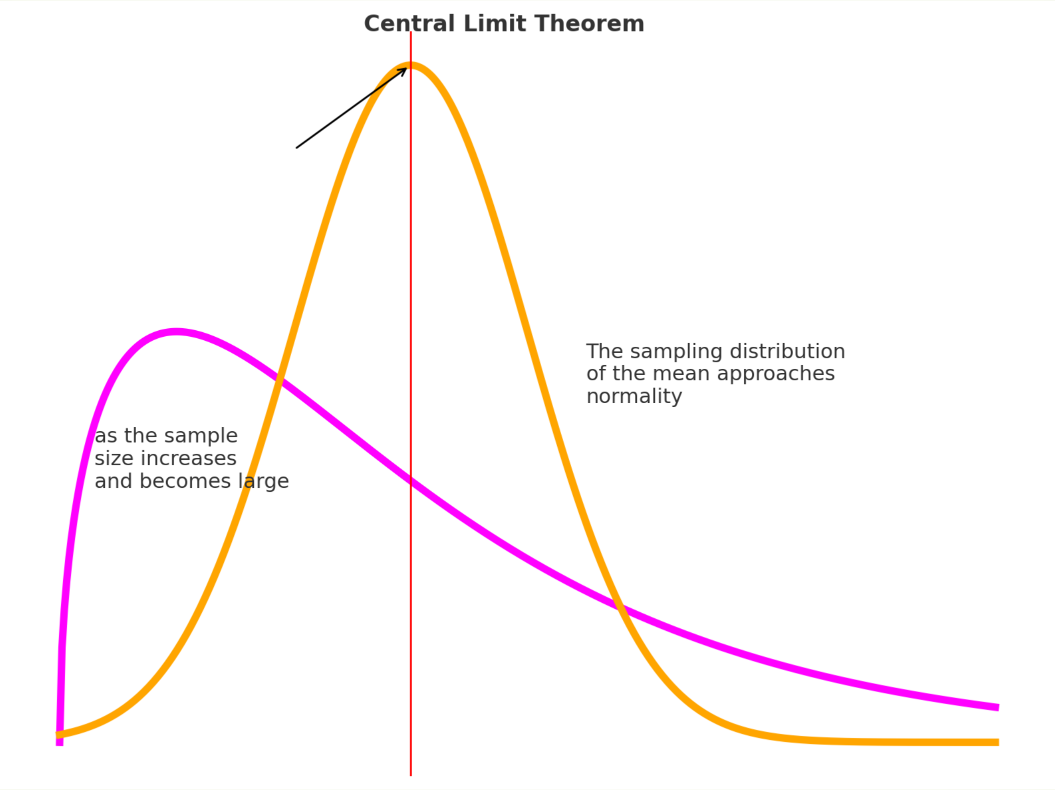
Sampling Distribution of Mean: Normal Population

- If $X_1, X_2, X_3, \ldots, X_n$ are n **independent random samples** drawn from a Normal Population with Mean = μ and Standard Deviation = σ, then the sampling distribution of $\bar{X}$ follows a Normal Distribution with Mean = μ, and Standard Deviation = σ/√n
- σ/√n is known by the term **Standard Error**

Central Limit Theorem

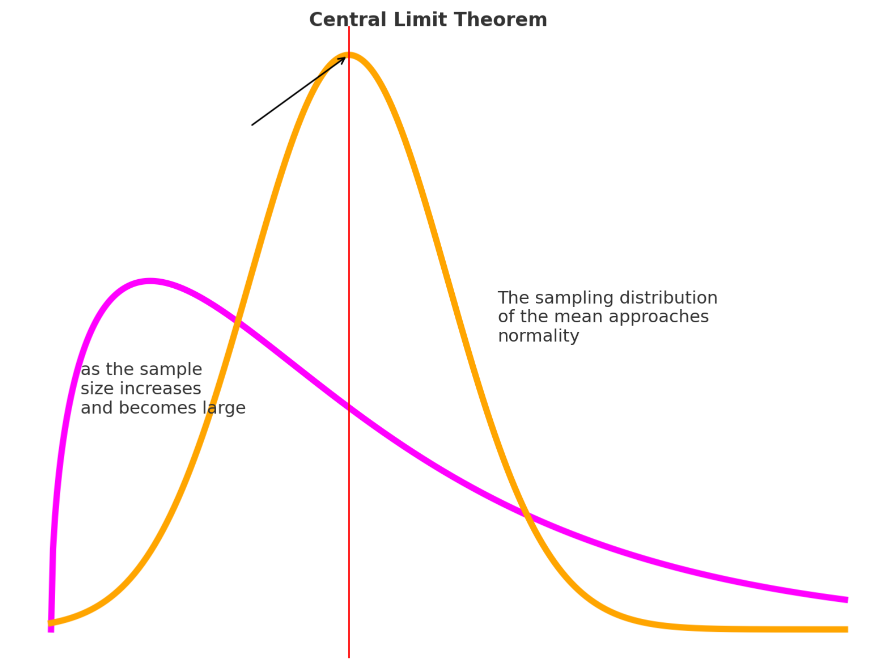
- *The distinguishing and unique feature of the central limit theorem is that* ***irrespective of the shape of the distribution of the original population****, the sampling distribution of the mean will approach a* ***normal distribution*** *as the size of the sample increases and becomes large*

CLT


Central Limit Theorem
as the sample
size increases
and becomes large
The sampling distribution
of the mean approaches
normality

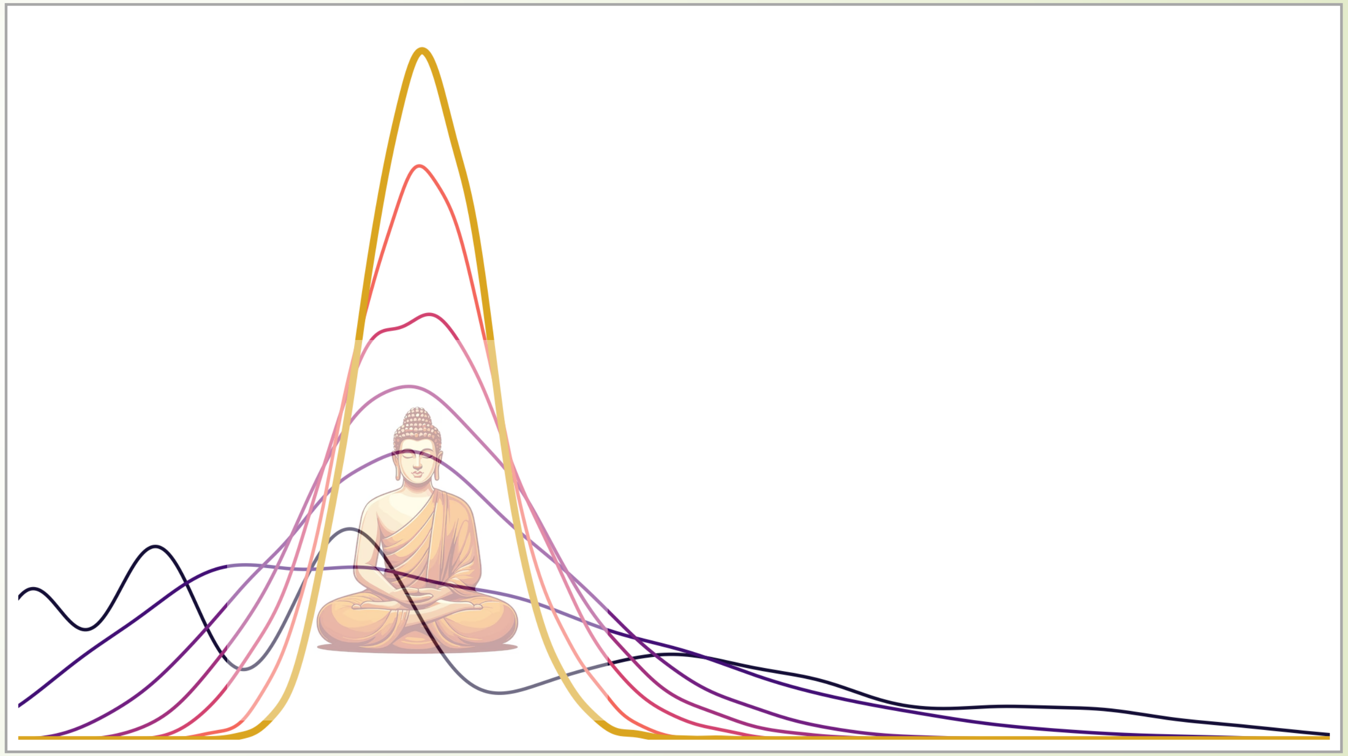
How large is large ?

- The Central Limit Theorem guarantees normality as n → ∞, but it does not specify how large n must be.
- The number 30 is a rule of thumb, not a mathematical cutoff.
- Statisticians tested the CLT across many underlying distributions (skewed, heavy-tailed, multimodal).
  - Found that by around n ≈ 30, the sampling distribution of the mean looked “close enough” to normal for practical purposes.
  - This benchmark came from simulation studies and practical experience, not a closed-form formula.
- Heavier skew and heavier tails may require larger n (sometimes 50, 100, or more).
- Symmetric or nearly normal populations need much smaller n (even n = 5 or 10 can suffice).

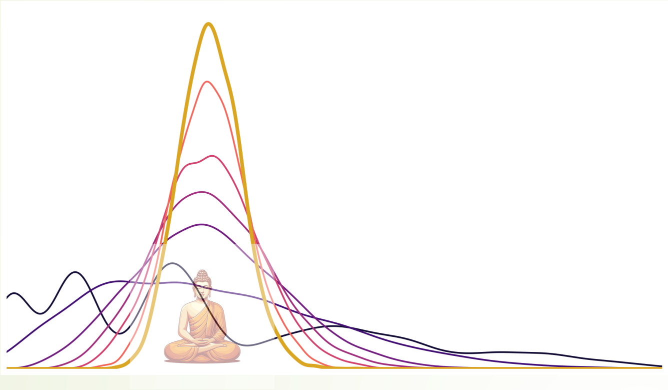

Central Limit Theorem
as the sample size increases and becomes large
The sampling distribution of the mean approaches normality

How does this work out ?

- Each sample mean is a sum of random variables divided by n.
- By **linearity of expectation**, the mean of sample means equals the population mean μ.
- The variance of the sample mean shrinks with sample size: Var($\bar{X}$) = σ^2/n.
- As n grows, repeated summing and averaging **smooths out irregularities** in the original distribution.
- **Convolution of distributions** (from adding random variables) drives the shape toward a bell curve.
- Result: $\bar{X}$ approaches N(μ, σ^2/n) as n becomes large.



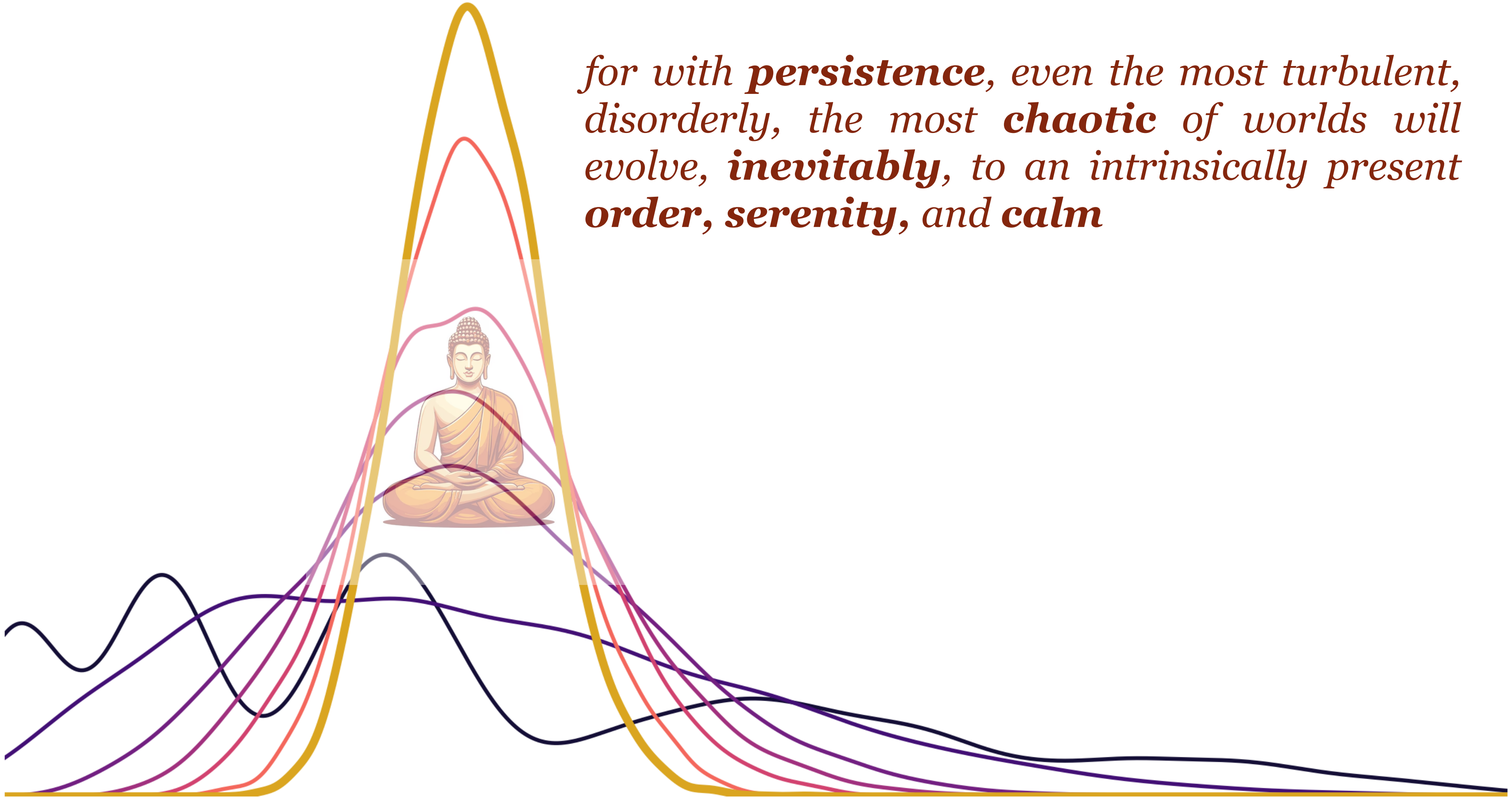
Convolution/Moving Windows



- In the CLT, convolution means adding random variables so their distributions combine and smooth into Normal; this is different from KDE's moving windows, which smooth data locally with kernels in x-space.
- Key difference:
  - KDE moving windows happen in **data space** (local smoothing of observed points)
  - CLT convolution happens in **distribution space** (repeated sums of random variables)

The Central Limit Theorem

simply, a convolution in distribution space



*for with **persistence**, even the most turbulent, disorderly, the most **chaotic** of worlds will evolve, **inevitably**, to an intrinsically present **order, serenity,** and **calm***

CLT

- Is **about the sampling distribution**
- It is **NOT** (any) of
  - The population distribution !
    - Rather, it characterizes the *distribution of averages across repeated samples*, not individual observations.
  - The PDF (of the raw data)
    - The CLT curve applies to sample means, not the underlying data points.
  - The empirical distribution of a dataset
    - It represents a theoretical distribution derived from repeated sampling.

What is the Utility Here

- We can use a relatively small sample to say something reliable about the whole population's average.
- We can attach a measure of how much uncertainty is in that estimate (the spread shrinks as the sample gets bigger).
- We can compare two groups
  - For instance, average income in one city vs. another, or average test scores between two classes.
- We can check if an observed average is plausible or unusual compared to what we would expect under "normal" variation.

Public Health: Daily Steps

- **Question:** Do users of the new fitness app walk more on average than non-users?
- The population data: daily steps are count data, skewed and often overdispersed (Negative Binomial), with many zeros.
  - Distribution: lumpy and skewed.
- But, the mean steps/day is approximately Normal.
- This lets us **make population-level inferences** about the app's effect on activity

Education: Tutoring Impact

- **Question:** Does after-school tutoring improve average test scores more than no tutoring?
- Data: score changes can be spiky near zero, asymmetric, with outliers (big gains or drops).
  - Distribution: asymmetric, sometimes bimodal (some benefit, some don’t).
- But, the mean gain is approximately Normal
- This enables **meaningful comparison** of program vs. control groups.

Genetics: Gene Expression

- **Question:** Does treatment X increase average expression of gene Y compared to control?

- Data: expression levels vary multiplicatively and are strongly right-skewed
  - Distribution: often log-Normal (microarray) or Negative Binomial (RNA-seq counts)
- But, the mean (or mean of log-values) is approximately Normal

- This lets us **test whether treatment alters expression in the population**, not just in noisy individuals

Nutrition: Dark Chocolate Benefits

- **Question:** Does eating dark chocolate lower blood pressure more than milk chocolate?

- Data: change in systolic BP (mmHg) is skewed with outliers (some large drops, some increases).
  - Distribution: right-skewed, heavy-tailed; sometimes a cluster near zero (non-responders)

- But, the mean change is approximately Normal

- This lets us **compare average effects between groups**
  - Decide if dark chocolate indeed helps more

Neuroscience/Medicine

- **Question**: Does a new memory training program improve average recall scores in early-stage dementia patients?

- Data: recall scores after intervention are bounded (0–100), clumped at low values, with a long right tail for patients who improve a lot.
  - Distribution: often skewed, heavy-tailed, and non-Normal.

- But, the mean recall improvement has an approximately Normal sampling distribution

- This allows researchers to **compare the average effect of the program against standard care**, despite noisy and uneven individual outcomes.

We are now knocking at the doors of **Hypothesis Testing**

More next week !