Stat C131A: Statistical Methods for Data Science

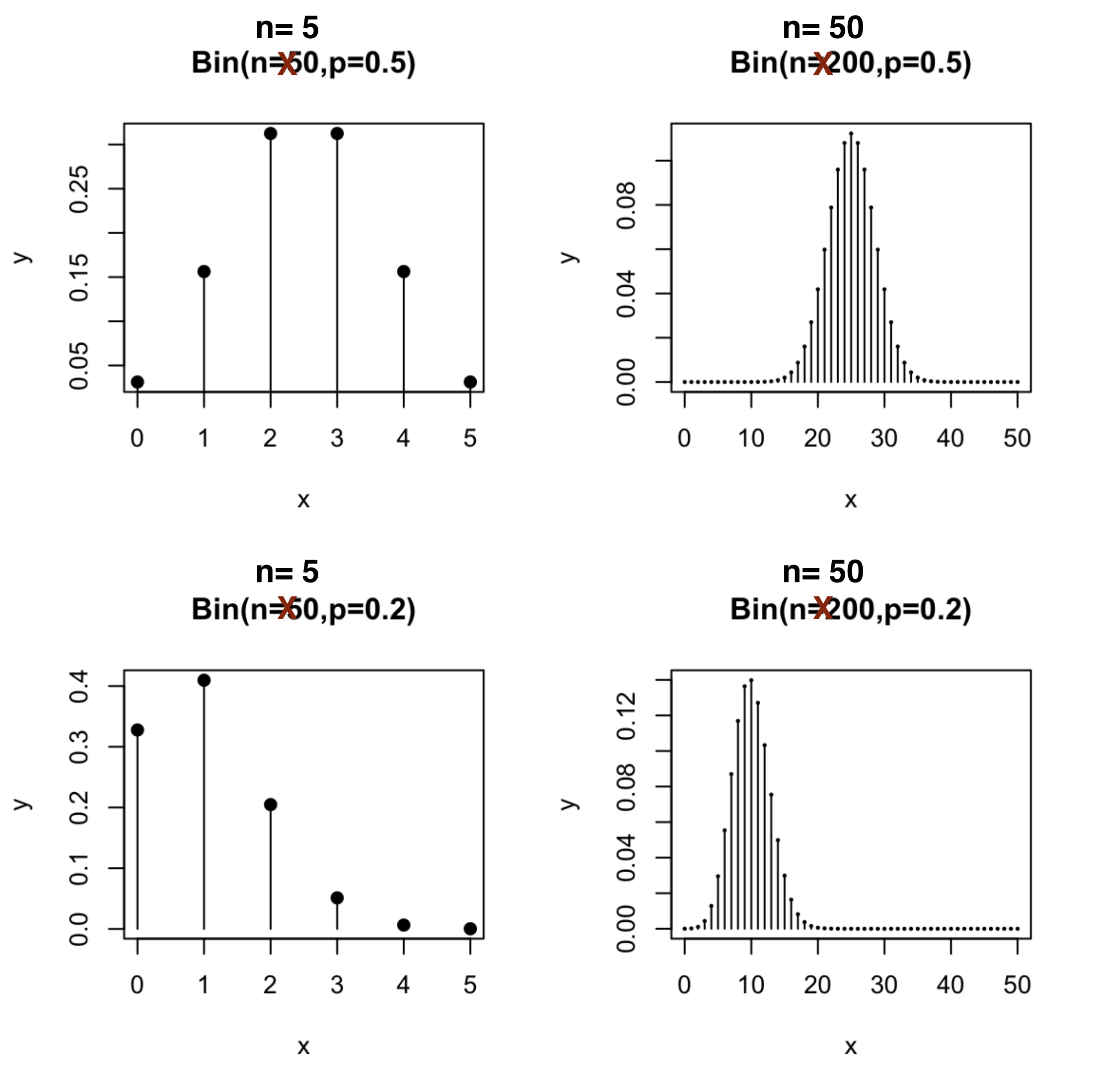
Lecture 6: Probability Distributions: Normal Distribution, Kernel Density Estimation

Sep 16 2025

Today

- Normal Distribution
- Density Curve Estimation

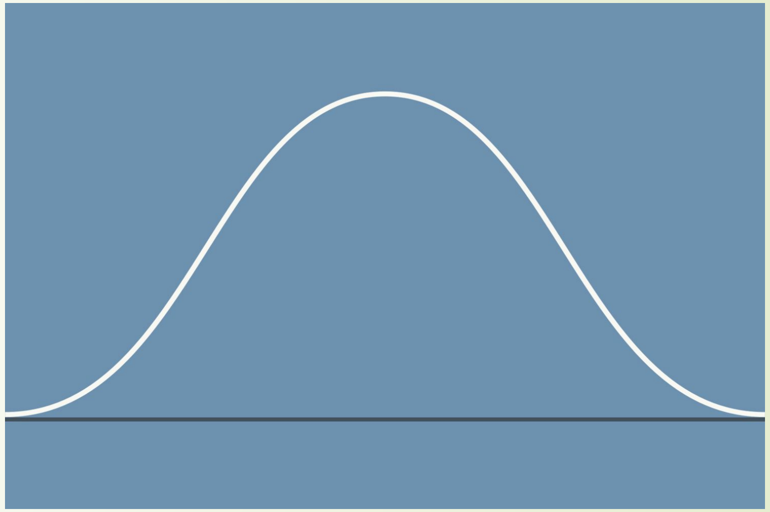
Error in panel



- The **peak** in the binomial distribution is at **np**

Bell Curve

- Symmetric, mound-shaped distribution of a continuous variable.
- Single peak at the mean → unimodal.
- Tails taper off in both directions.
- Informally: any symmetric, bell-shaped density curve.
- The Normal distribution is the most important example.



Why are Bell Curves Popular

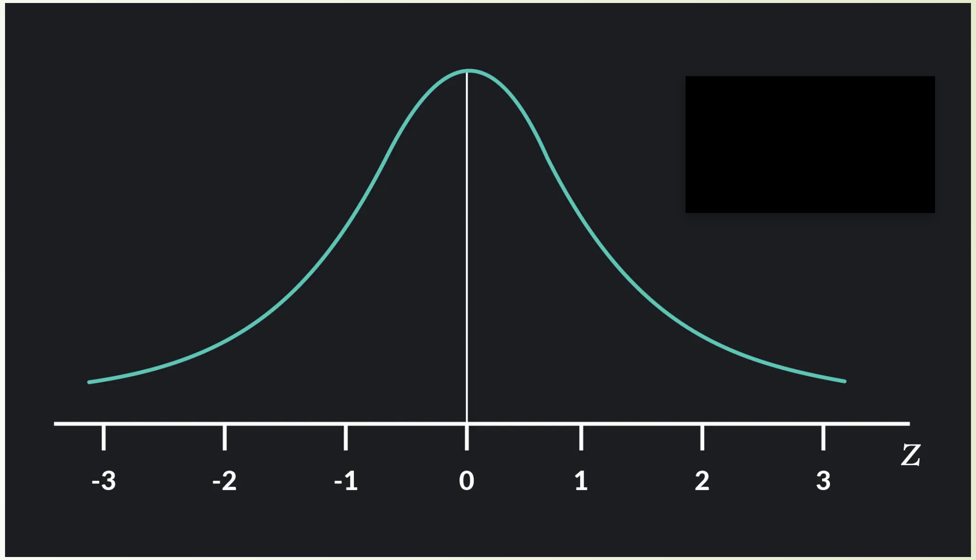
- Easy to recognize and interpret: most values cluster around the center, fewer at the extremes
- Symmetry makes the center a natural “typical” value
- Single peak highlights the most common outcome
- Gradual tapering of the tails matches the idea of rare extremes
- Often a good first approximation for real-world data like test scores, heights, and errors
- Provides a simple baseline for comparing other shapes (e.g., bell-shaped but heavier tails)
- Communicates variation clearly to non-specialists : easy to explain “typical vs. rare”
- Shared by many different distributions (Normal, t, Laplace, Logistic), making it broadly useful across fields

The Normal Distribution

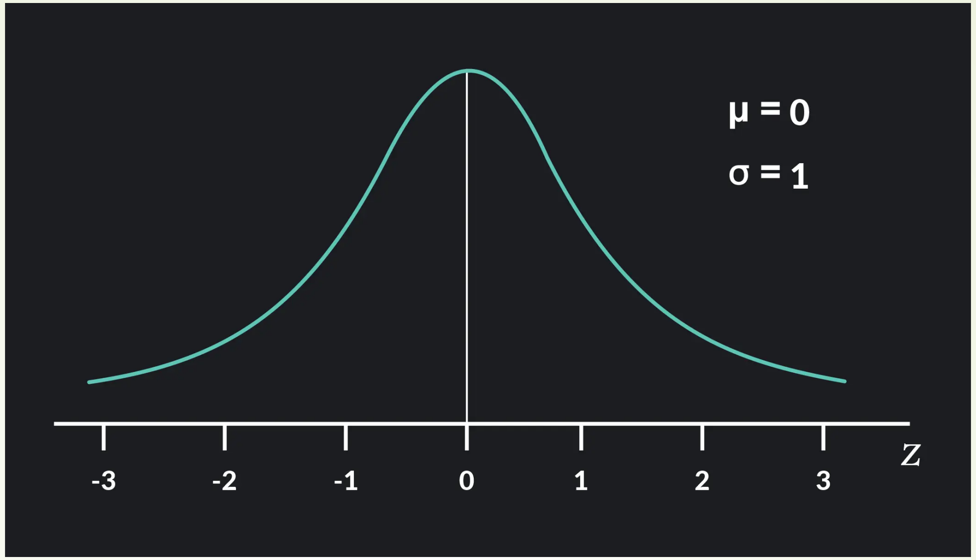
- Defined by two parameters:
  - **μ (mean):** location of the center.
  - **σ (standard deviation):** spread/width of the curve
- Defined by the **Gaussian** function

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

- **Properties:**
  - **Mean = Median = Mode.**
  - Total area under curve = 1.
  - Highest probability density at μ.

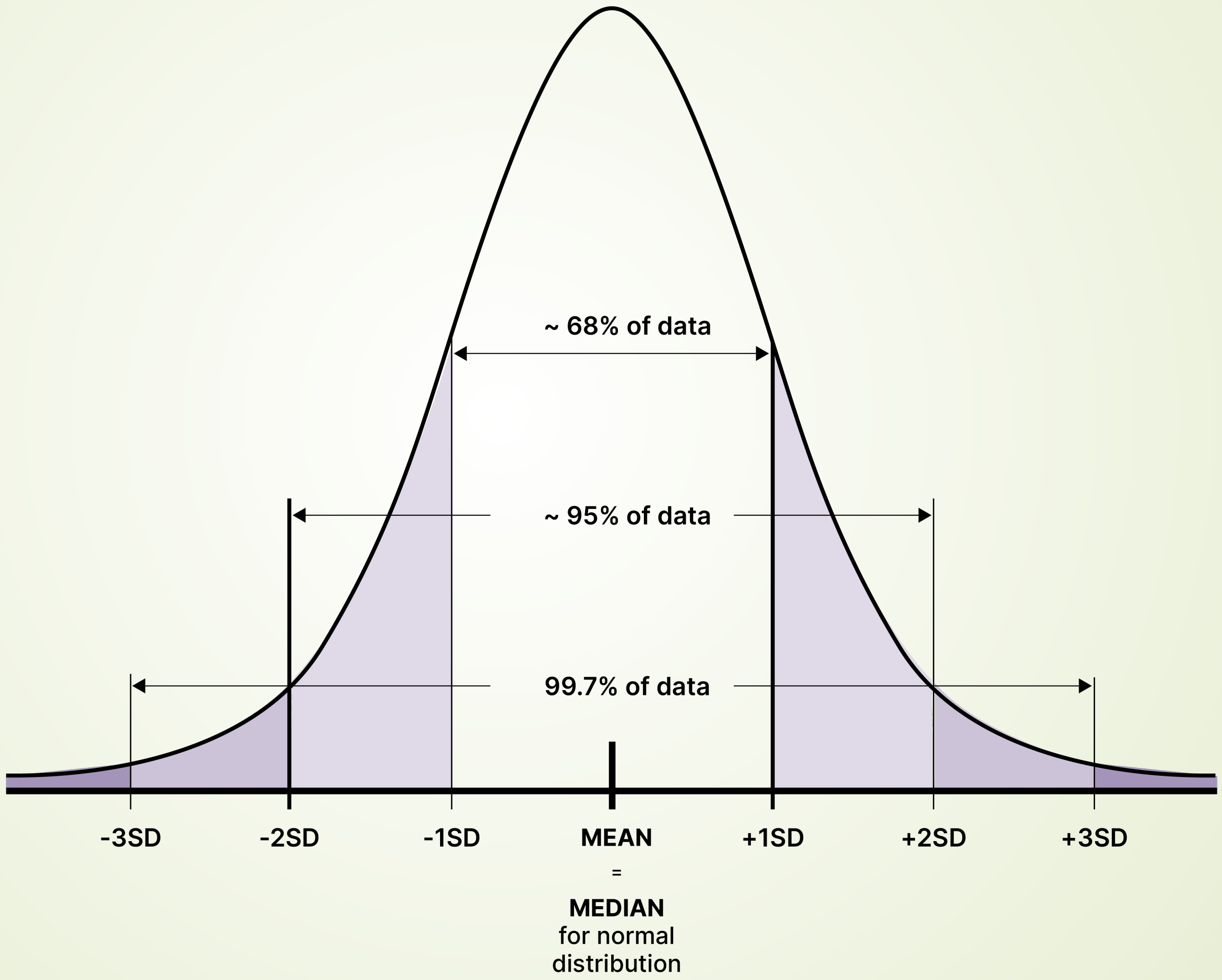


Attributes



- Bell-shaped, symmetric about μ.
- Often described as **"Beautiful"** **!**
- Probability measured as **area under curve**, not frequency counts.
- Larger $\sigma \rightarrow$ wider, flatter curve; smaller $\sigma \rightarrow$ narrower, taller curve.
- **Basis of much of statistical inference !!**
- The **standard normal distribution** has $\mu = 0$, $\sigma = 1$

Empirical Rule: (68–95–99.7 Rule)



The Z-Score

- **Definition:** A Z-score tells how many standard deviations a data point x is from the mean
- $z = (x - \mu) / \sigma$
- **Interpretation:**
  - Positive z → value is above the mean.
  - Negative z → value is below the mean.
  - z = 0 → value equals the mean.
- **Unit-free:** Z-scores remove original units (dollars, inches, seconds) and put all values on the same scale.
- **Comparability:** Enables comparison across different datasets with different means and spreads (e.g., test scores on different exams).
- **Probability link:** By converting to z, we can use the **standard Normal table** to compute probabilities for any Normal distribution.
- **Thresholds:**
  - $|z| \approx 2$ → "unusual" values (outside ~95% range).
  - $|z| > 3$ → very rare (outside ~99.7% range).
- **Applications:**
  - Outlier detection
  - Standardized test scoring
  - …
- Input to many statistical tests and confidence intervals.

Evolution of the Normal Distribution

- 1733: de Moivre approximates the Binomial distribution with a bell-shaped curve, published in *The Doctrine of Chances*
- 1809: Gauss uses the distribution to model astronomical measurement errors (*Theoria Motus*), leading to the name "Gaussian"
- 19th century: Quetelet and Galton promote the curve as a model for human and social traits, popularizing the term "Normal"
- 20th century onward: Normal distribution becomes foundational in statistical inference, hypothesis testing, and modern probability theory

Density Curve Estimation

Purdom Texbook: 2.5

Introduction

- Goal: Estimate underlying density curve of data distribution
- Would like an estimate of **the unknown pdf** $p(x)$ f**or the distribution that created the data**
- Density estimation provides smooth, interpretable functions
- **Density** Histograms: area underneath sums to 1

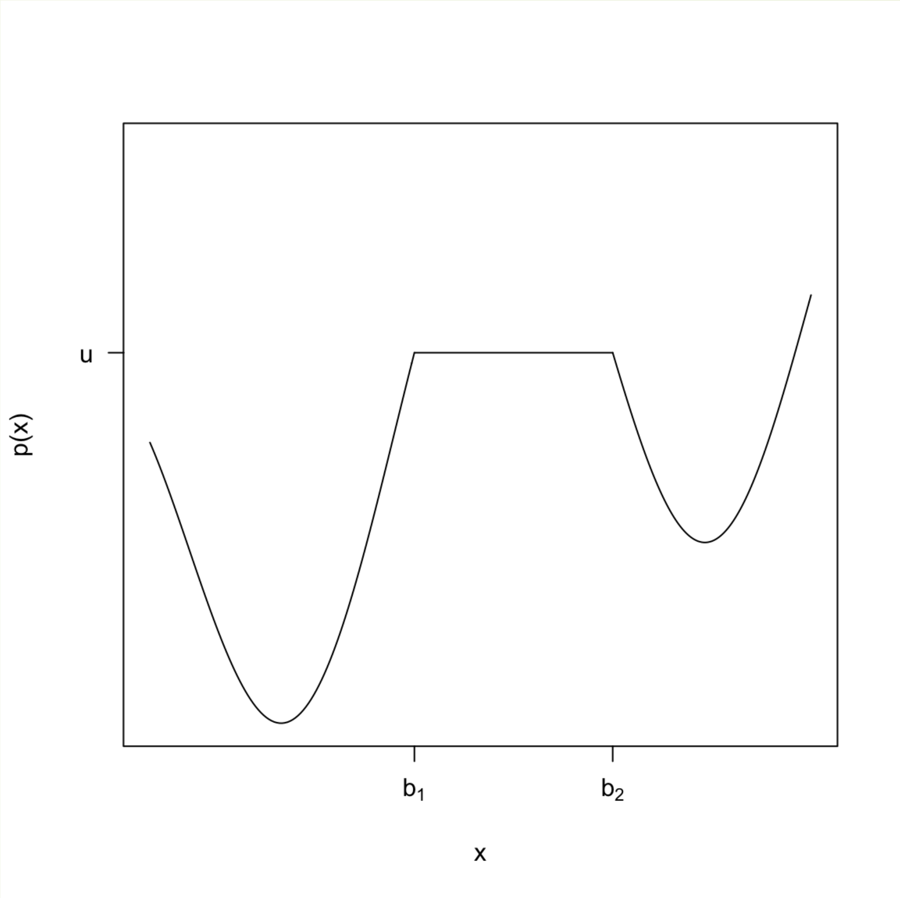
Histograms: Recap

- Divide range into bins of equal width
- Count number of observations in each bin
- Visualize as bars

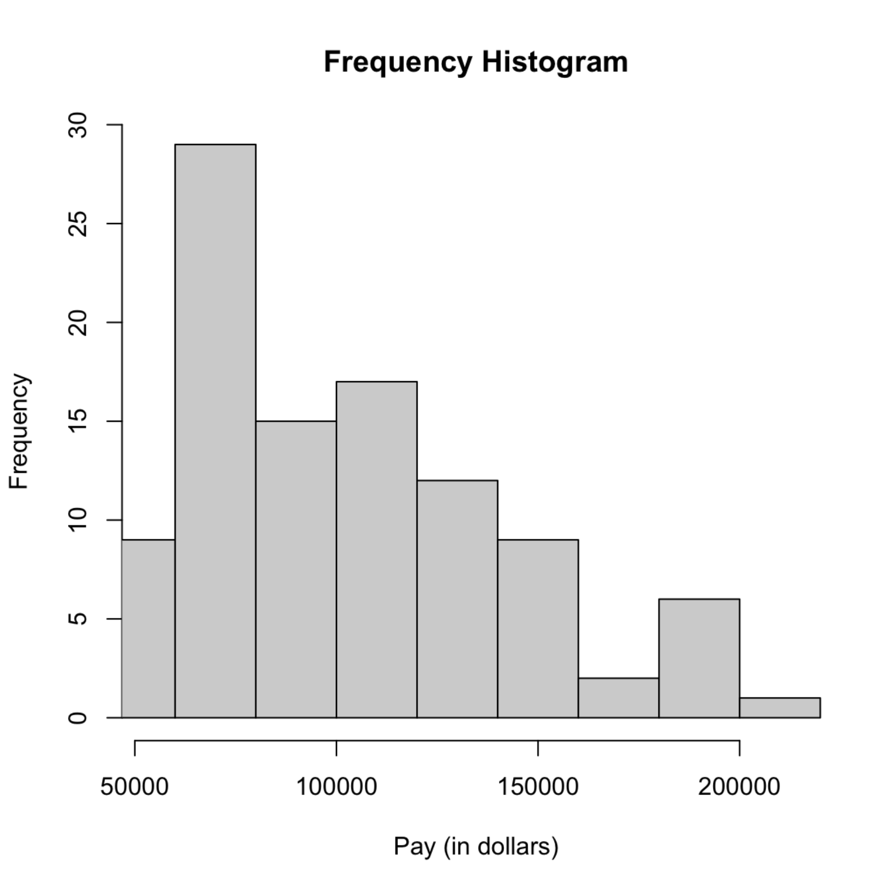
Density Histogram

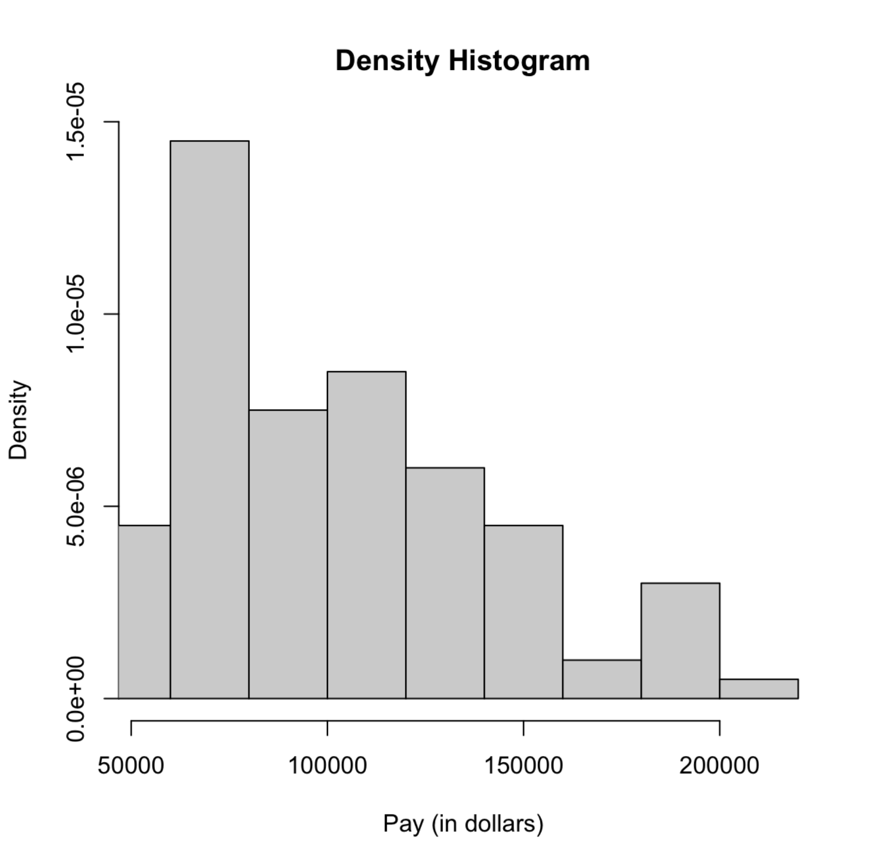
- **Density Histograms**
  - Y-axis scaled so total area = 1
  - Area of bars corresponds to probability
  - Formula: height = count / (n * bin width)

- **Probability**
  - $P(X \in [b_1, b_2]) = u * (b_2 - b_1)$
  - $\hat{P}(b_1 \leq X \leq b_2) = \dfrac{\text{\# Points in } [b_1, b_2]}{n}$
  - $\hat{p}(x) = \hat{P}(b_1 \leq X \leq b_2)/(b_2 - b_1) = \dfrac{\text{\# Points in } [b_1, b_2]}{(b_2 - b_1) \times n}$



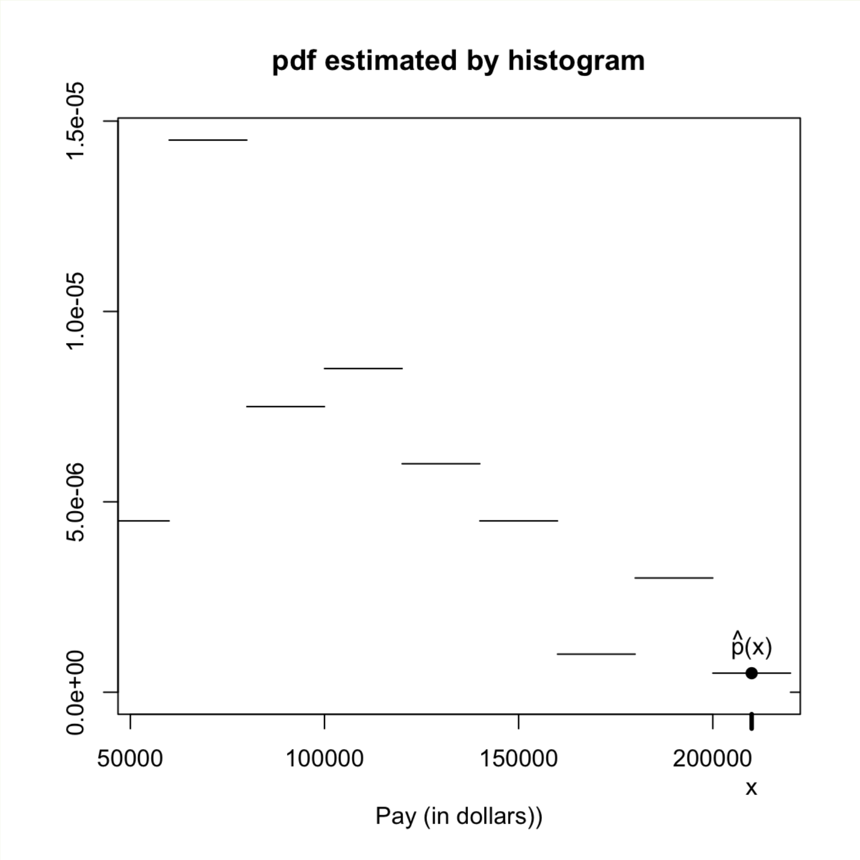
Revisiting Density Histograms





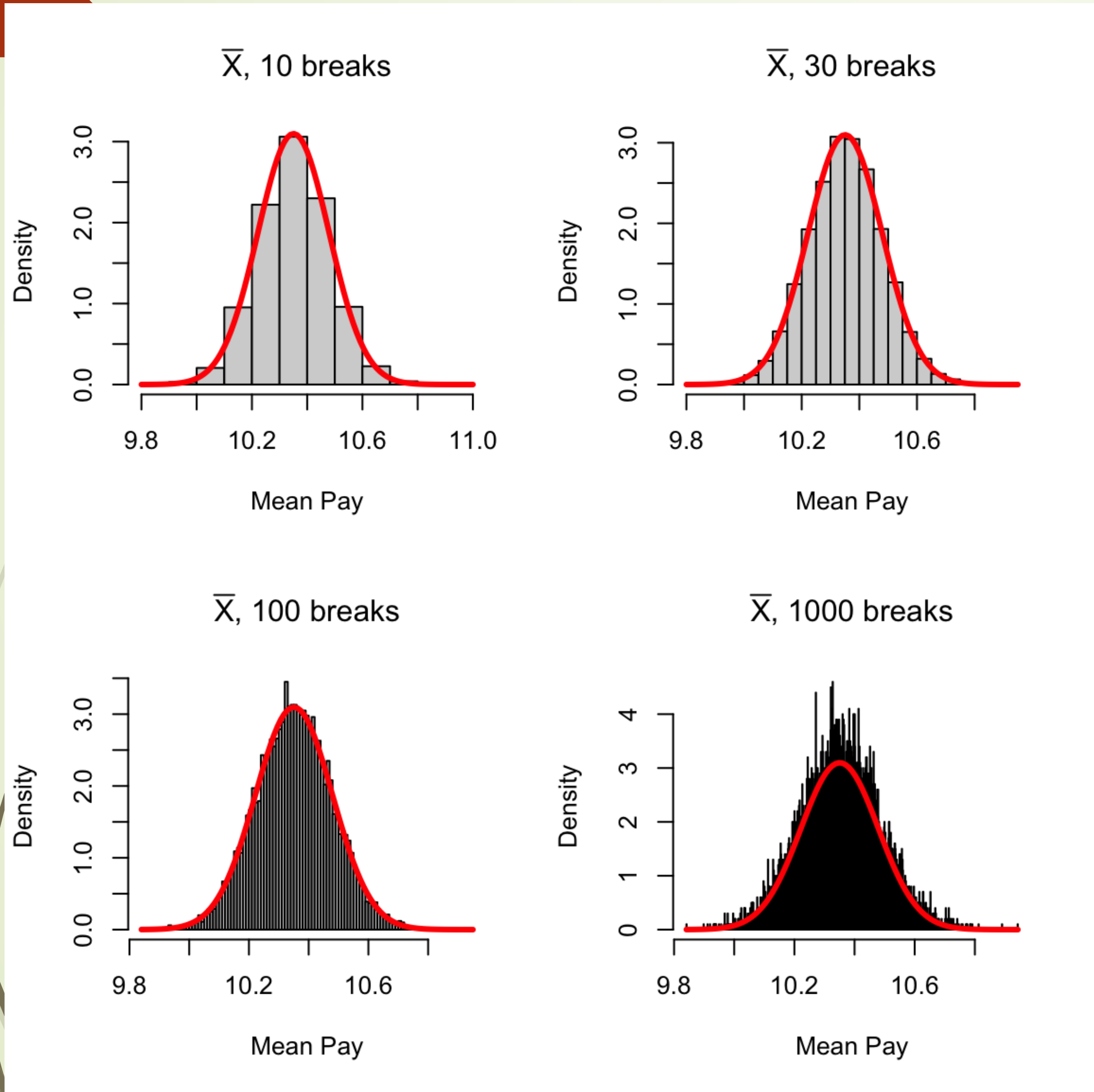
- $\dfrac{\text{\# Points in } [b_1, b_2]}{(b_2 - b_1) \times n}$

Step Function



- **Step** function: **PDF estimated by histogram**
- $$\hat{p}_{hist}(x) = \frac{\hat{P}(\text{data in bin of } x)}{w}$$

Limitations of Histograms



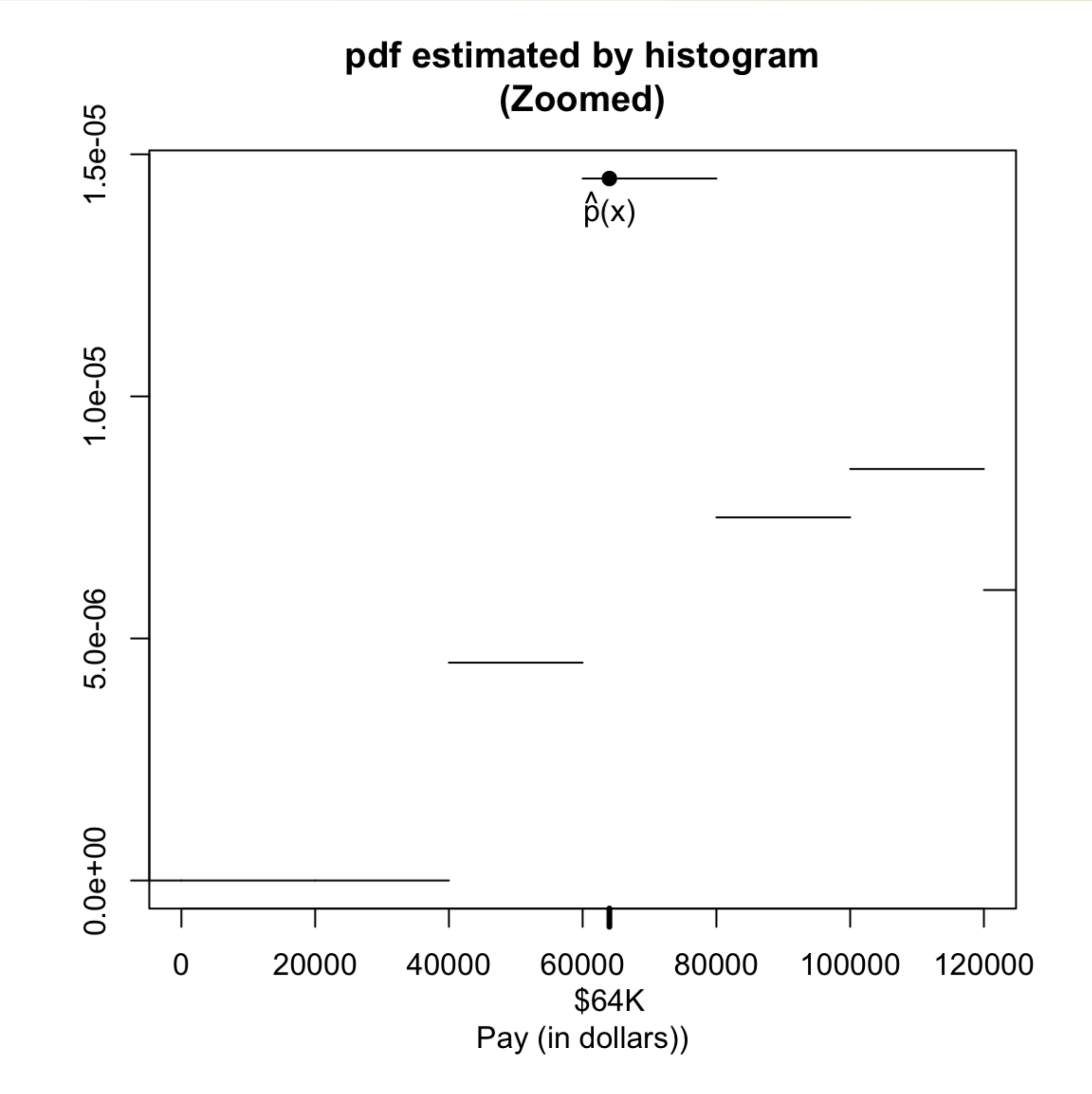
- **Choice of bin width** strongly affects appearance
- Discontinuous, blocky representation
  - Not smooth

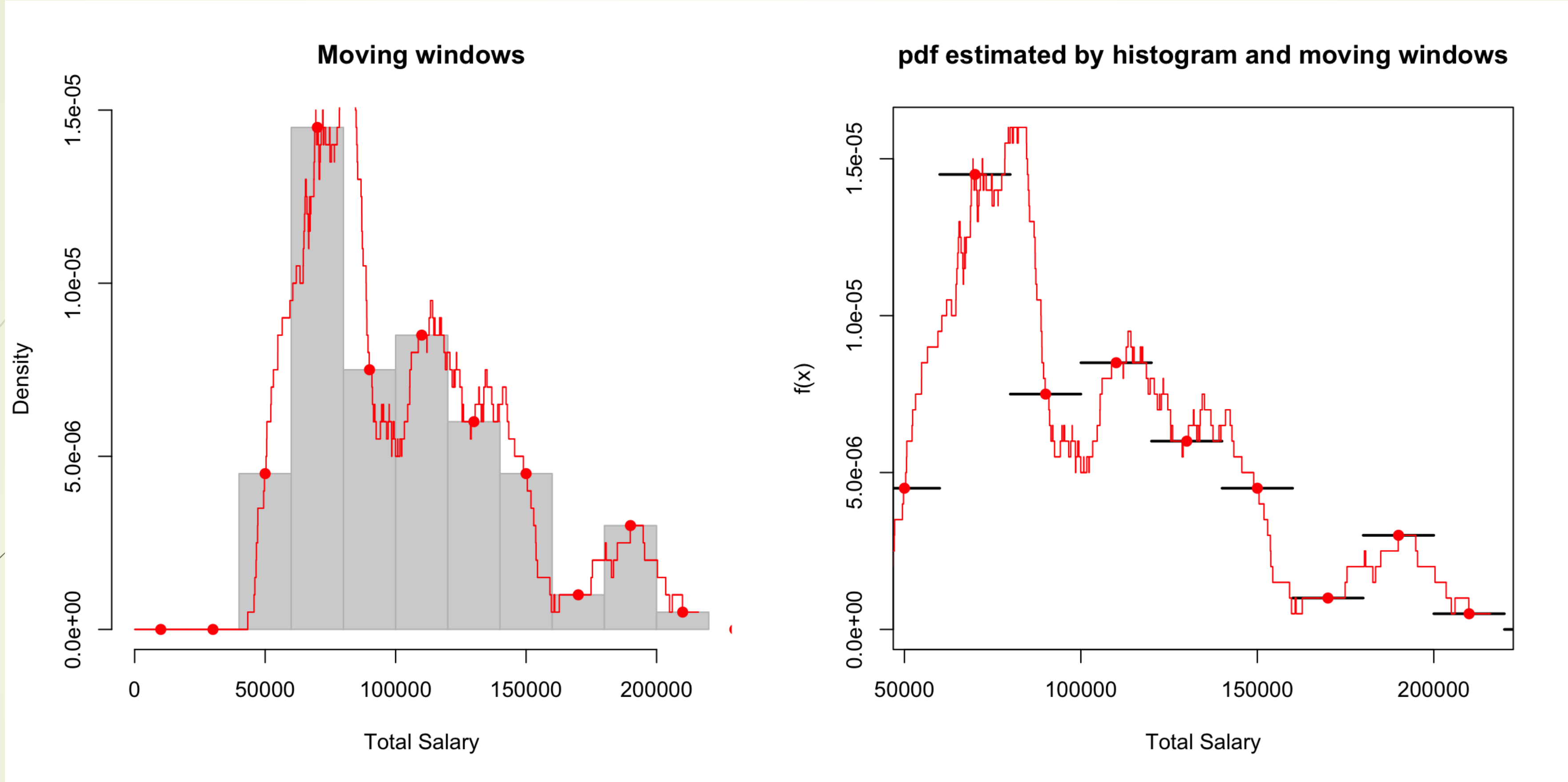
Kernel Density Estimation (KDE)

- Idea: place smooth bump (kernel) at each data point
- Sum all bumps to estimate density
- Produces continuous, smooth curve
- The curve is scaled so the total area = 1, just like a probability distribution

Moving Windows

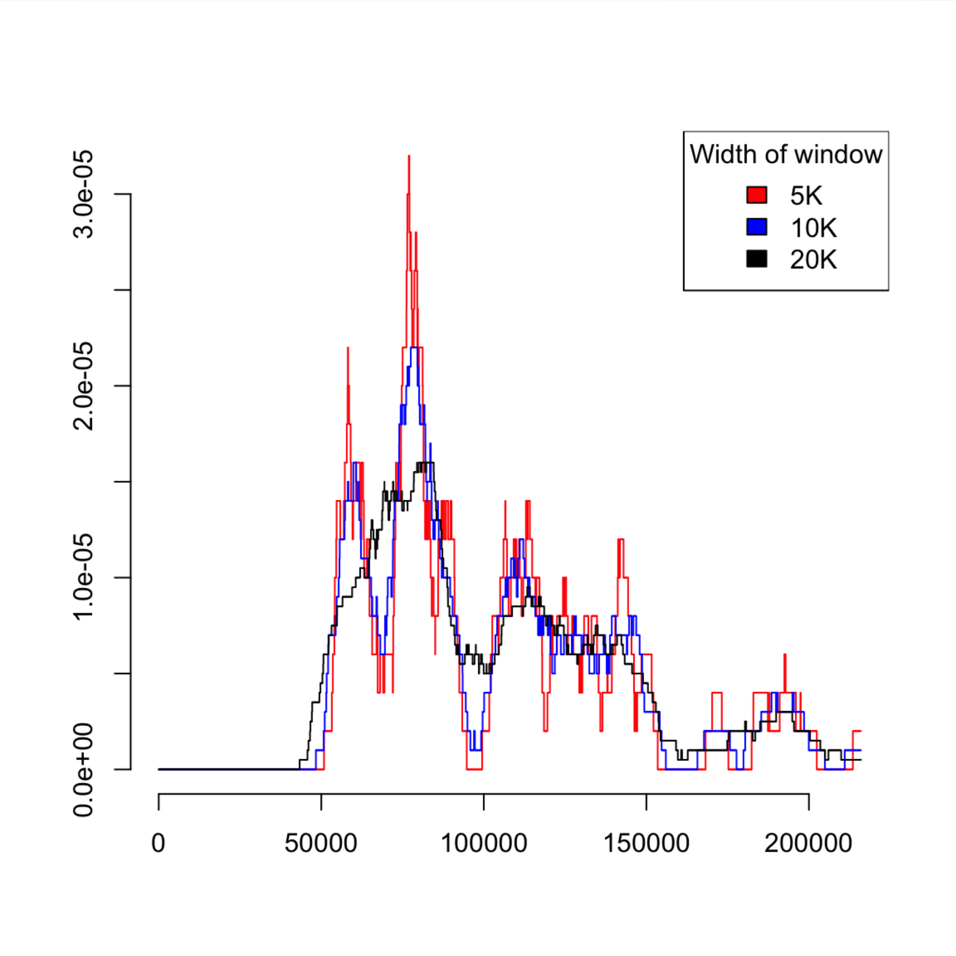
- **PDF**
  - Smooth function
  - Assume "flat" (wont change much) in small window
- For 64 K
  - $(b_1, b_2) = (54,000, 74,000)$
  - $\hat{p}(64,000) = \dfrac{\text{\# Points in } (b_1, b_2]}{(b_2 - b_1) \times n} = \dfrac{\text{\# Points in } (54K, 74K]}{20K \times n}$
  - $\hat{p}_{hist}(64,000) = \dfrac{\text{\# Points in } (60K, 80K]}{20K \times n}.$
- Not an ideal way to do this !
- **Convolution**





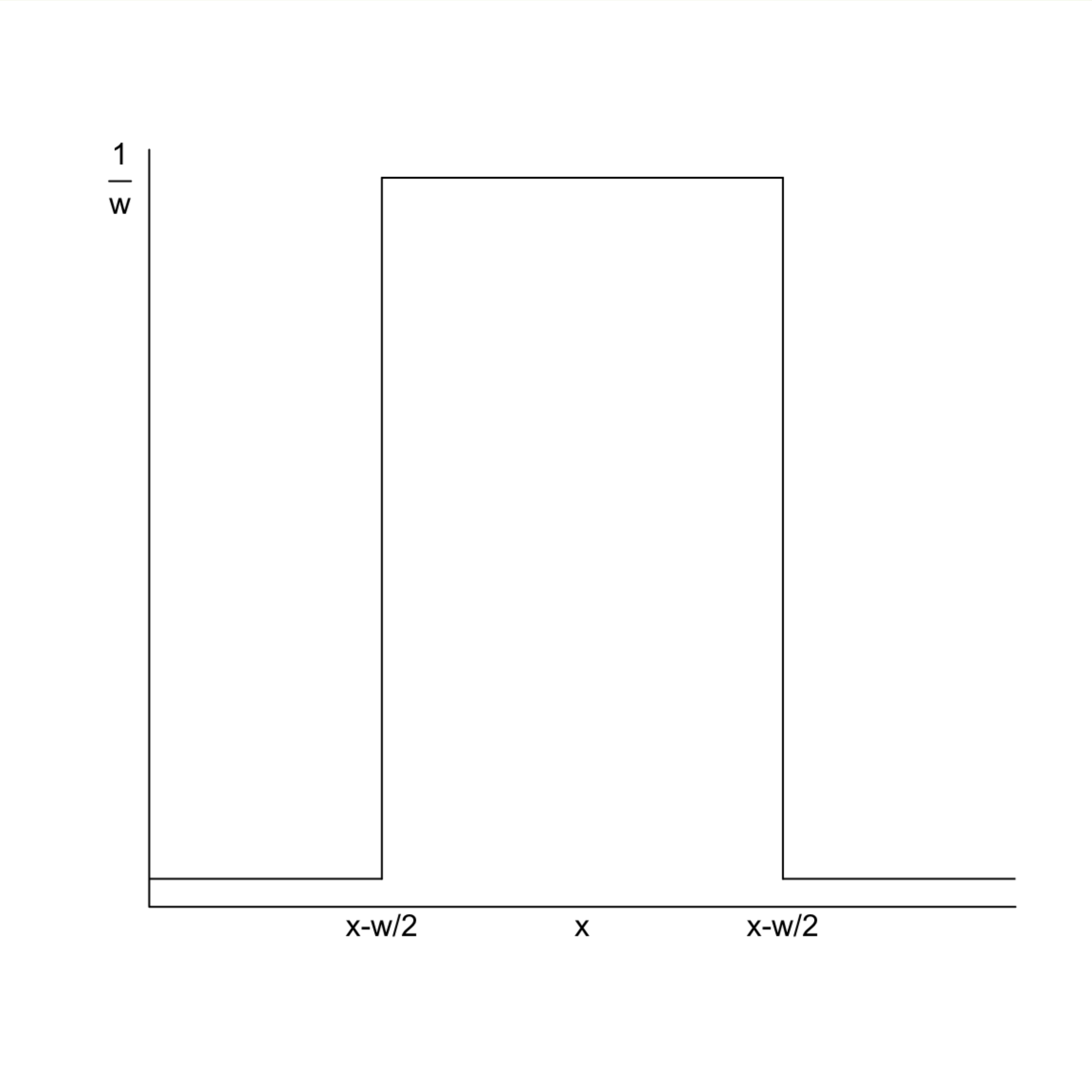
- $\hat{p}(x) = \dfrac{\#X_i \in [x - \frac{w}{2}, x + \frac{w}{2})}{w \times n}$
- w is the window size

Different size windows



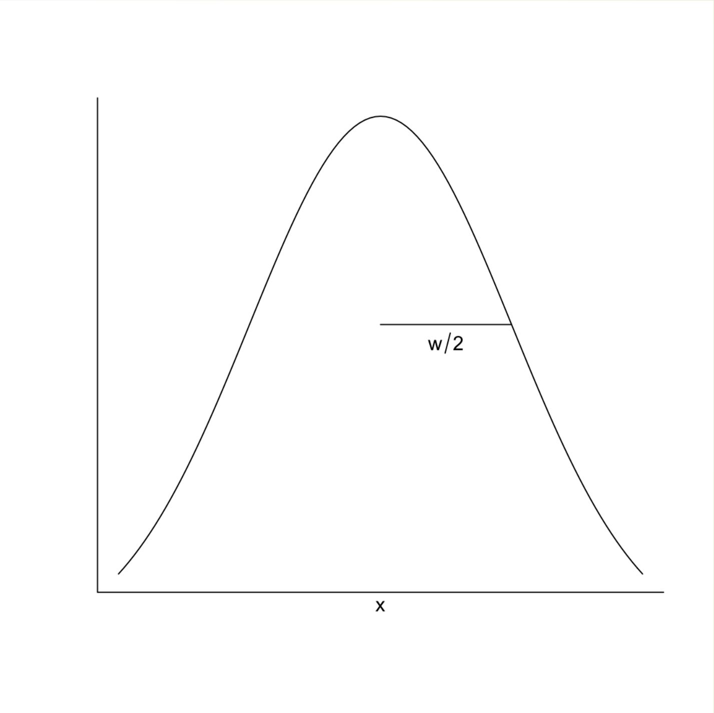
- Effect of window size ?

Boxcar Kernel

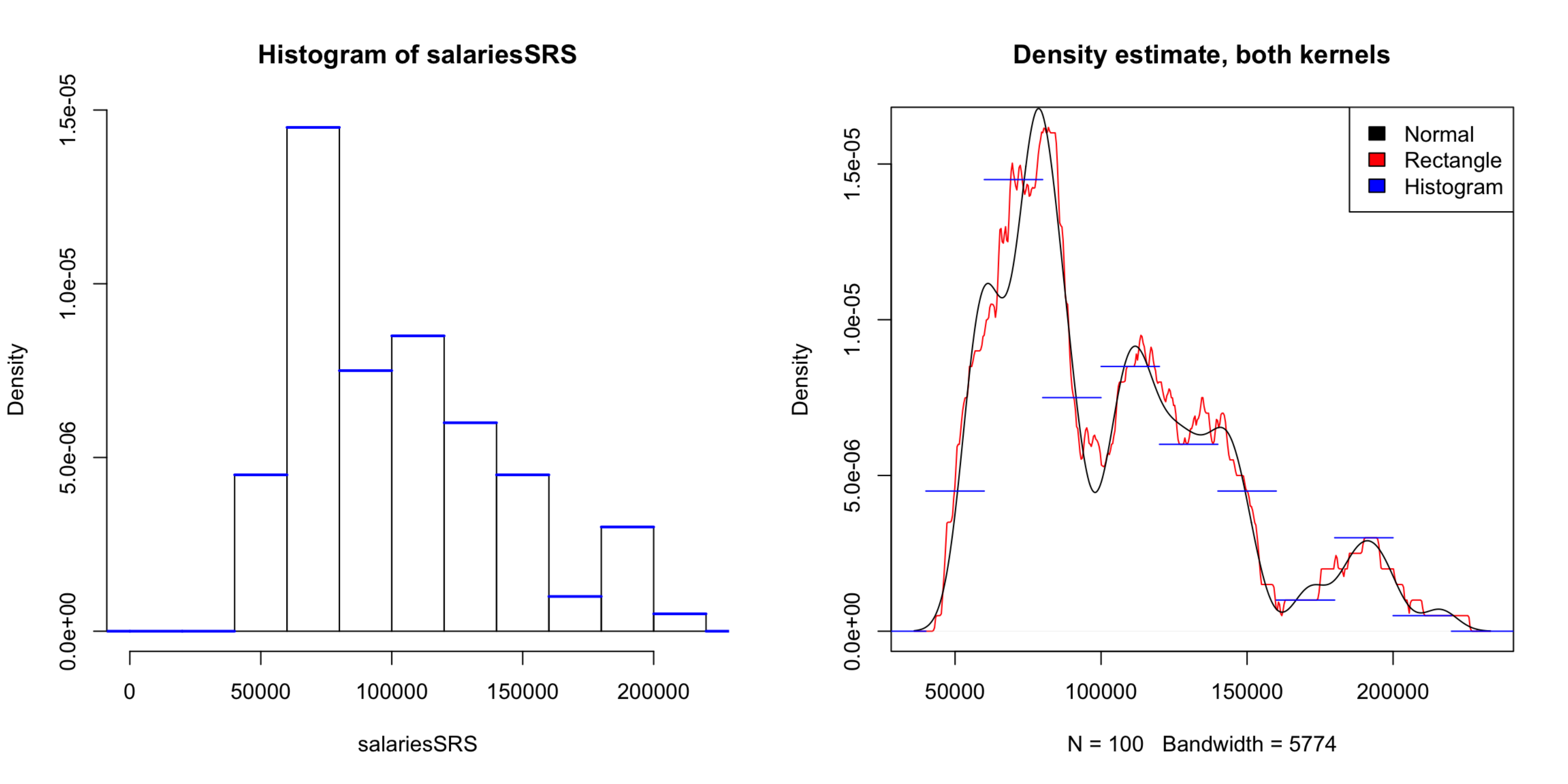


Normal (distribution) Kernel

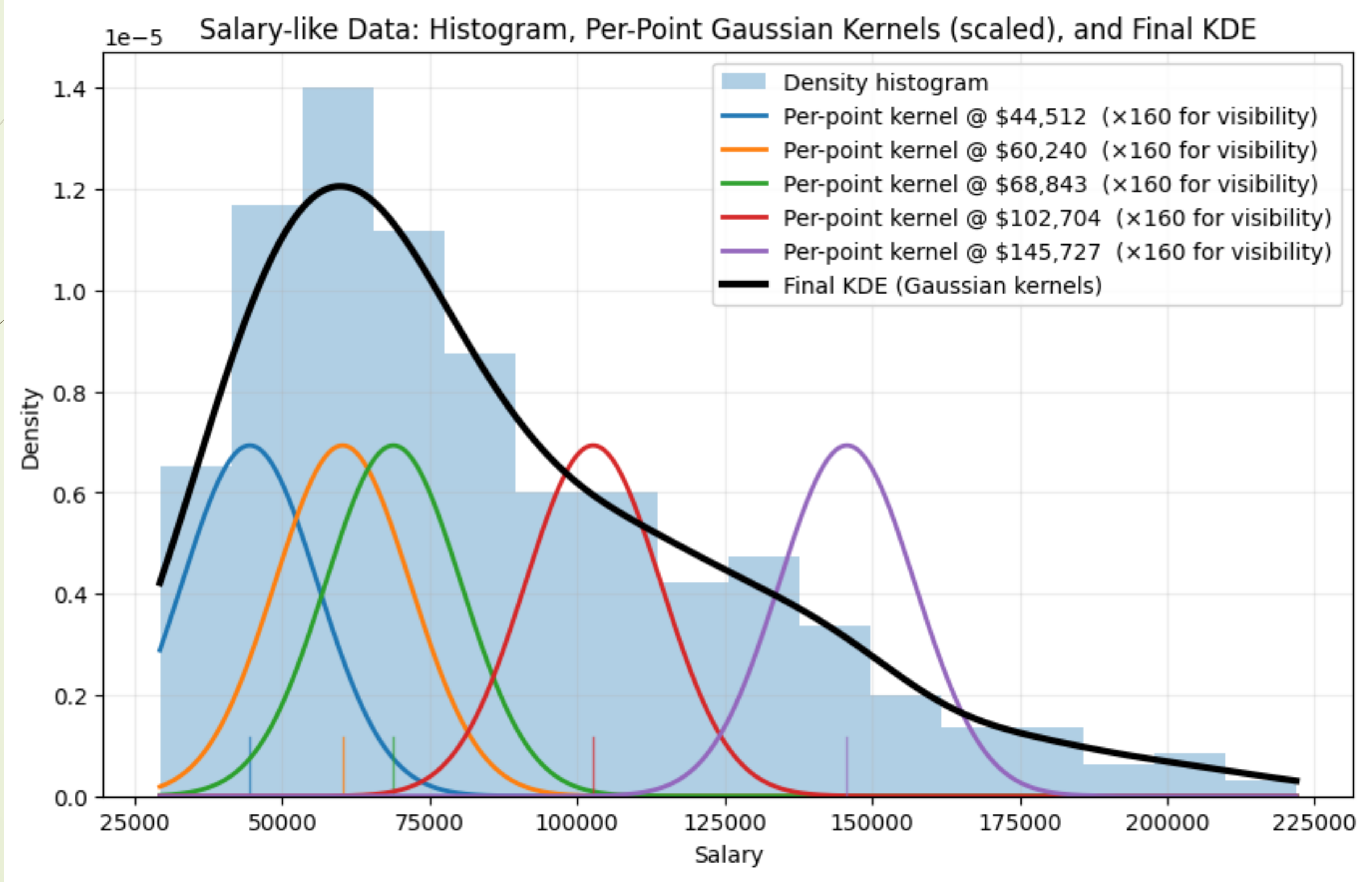
- h is the **bandwith parameter**
- $\frac{1}{h}K(\frac{|x - X_i|}{h})$
- Normal curve, with
  - Mean = x
  - Standard deviation = h



Histogram, Rectangle, Normal



Sum (average !) of all Curves/Bumps



Choice of Kernel

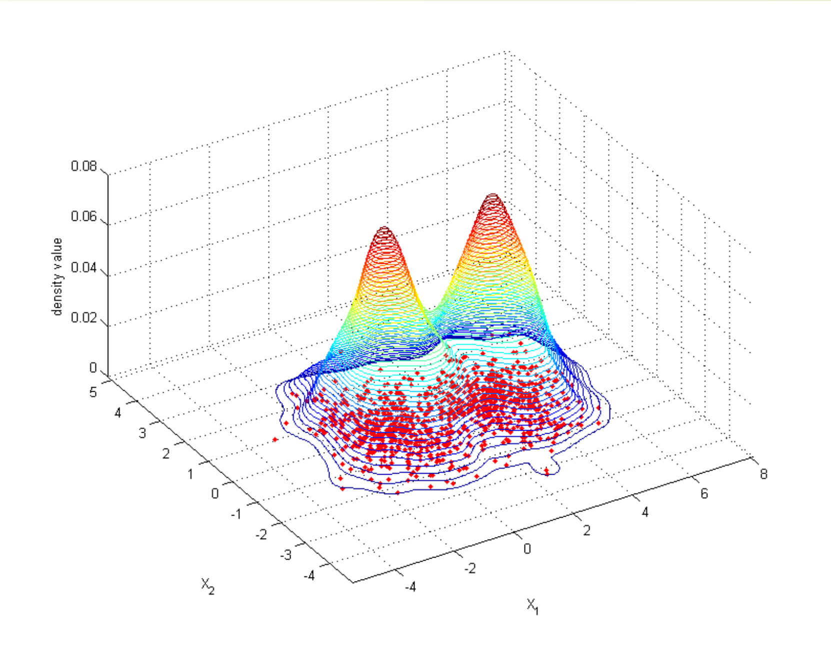
- Common kernels: Gaussian, Epanechnikov, Uniform
- All valid as long as integrate to 1
- Shape less important than bandwidth

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



Comparing Groups with Density Curves