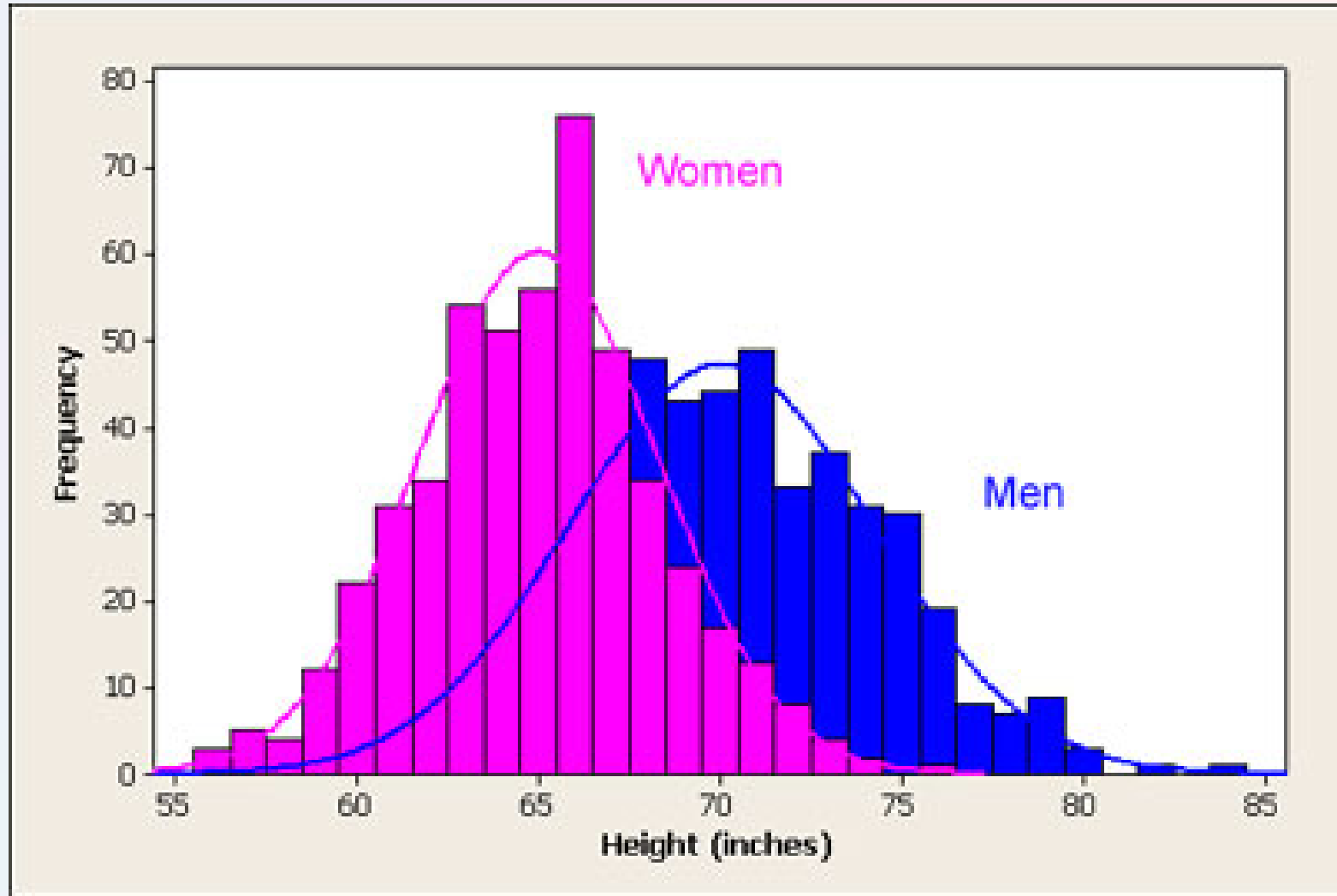


# Density Estimation

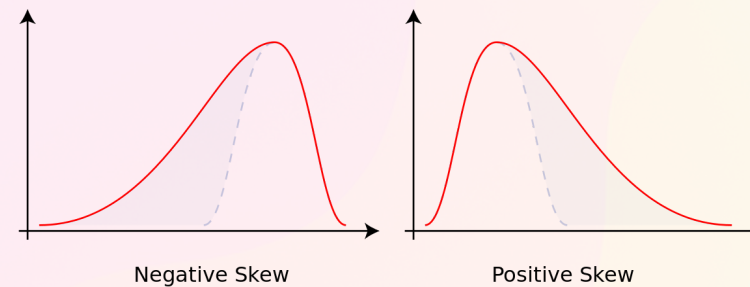
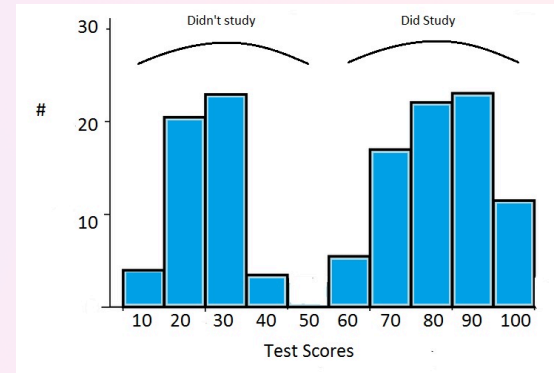
David I. Inouye

# Density Estimation Finds a Density (PDF/PMF) That Represents the Data (or Empirical Distribution) Well



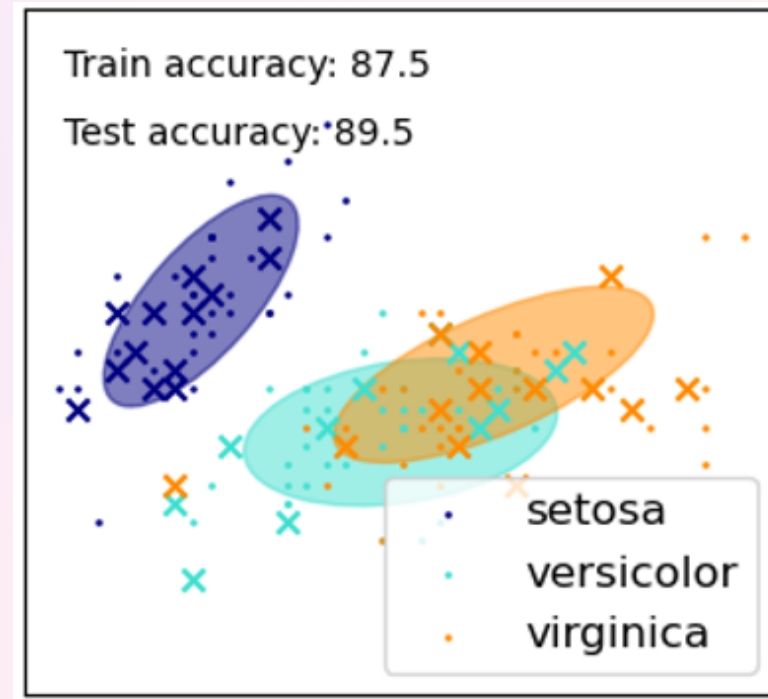
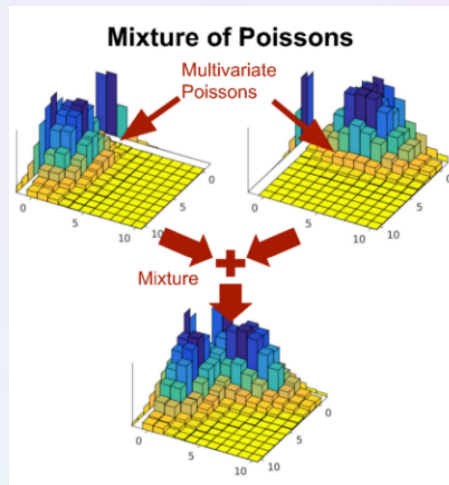
# Motivation: Density Estimation Can Be Used to Uncover Underlying Structure

- Uncover multi-modal structure
- Uncover skewness



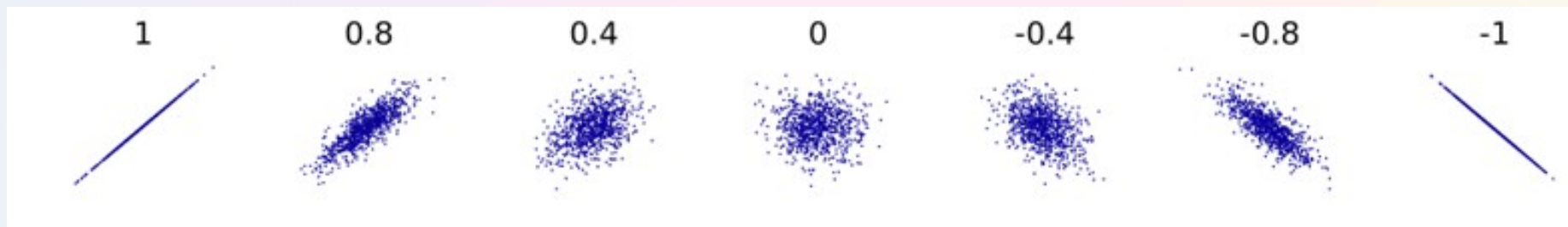
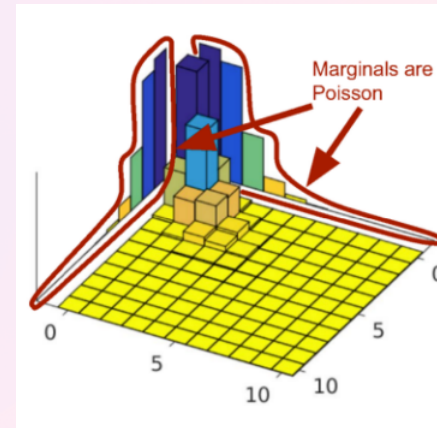
# Motivation: Density Estimation Can Be Used to Uncover Underlying Structure

- Cluster structure
  - Gaussian mixture models
  - Poisson mixture models

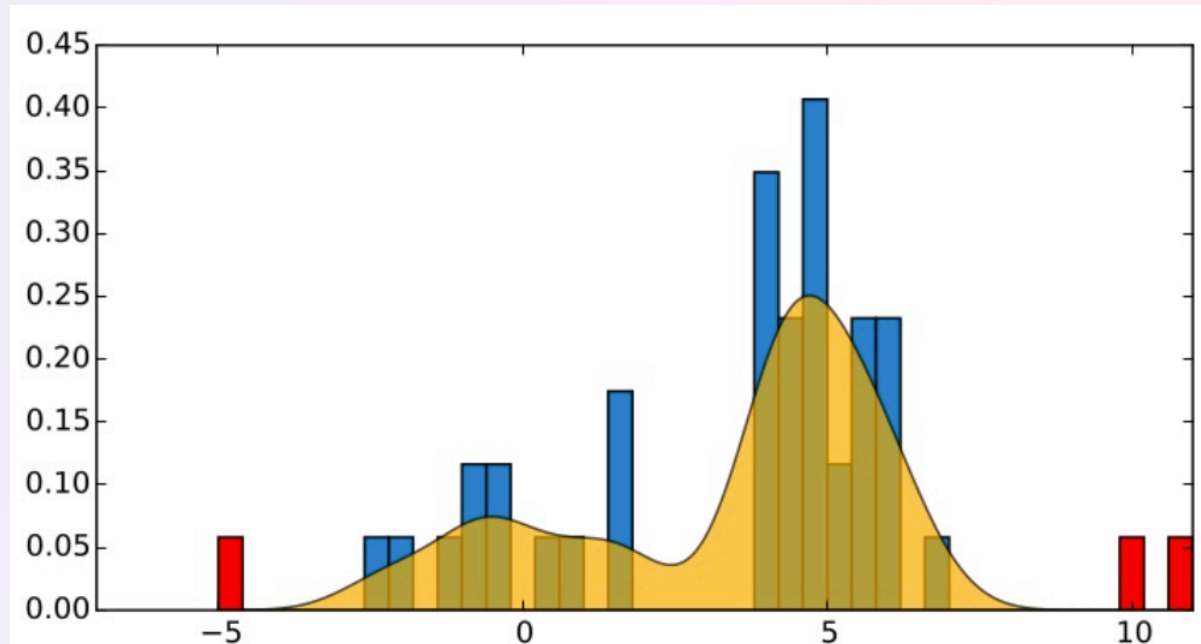


# Motivation: Density Estimation Can Be Used to Uncover Underlying Structure

- Dependence structure of random variables (e.g., correlation)



# Motivation: Density Estimation Can Be Used for Anomaly Detection



# Parametric Density Estimation Assumes a Density Model Class Parameterized by $\theta$

- **Assumption: Bernoulli density**
  - $\theta = [p], \quad p \in [0, 1]$
- **Assumption: Exponential density**
  - $\theta = [\lambda], \quad \lambda \in \mathbb{R}_{++}$
- **Assumption: Gaussian density**
  - $\theta = [\mu, \sigma^2], \quad \mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}_{++}$
- **Assumption: DNN-based model**
  - $\theta = [\text{“all neural network parameters”}]$

# How Do We Determine Which Model in the Model Class Is the Best?

- Classically, people have turned to information theoretic quantities
  - Entropy
  - Kullback Liebler (KL) Divergence
  - Maximum likelihood estimation (MLE)

# Informally, **Entropy** Measures the “Amount of Randomness/Disorder” of a Distribution

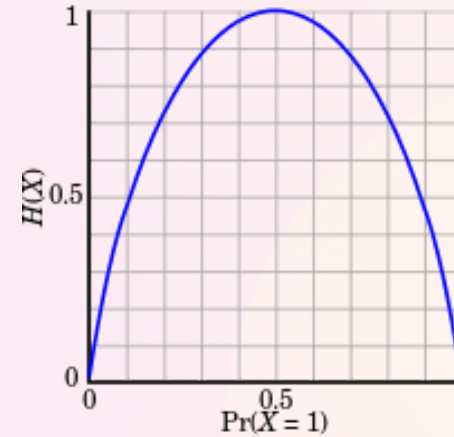
- Formally, **entropy** for discrete variables:

$$H(P(\cdot)) = \mathbb{E}[-\log P(x)] = - \sum_x P(x) \log P(x)$$

- Formally, **differential entropy** for continuous variables:

$$H(p(\cdot)) = \mathbb{E}[-\log p(x)] = - \int_x p(x) \log p(x) dx$$

- Consider fair coin vs coin where both sides are heads.



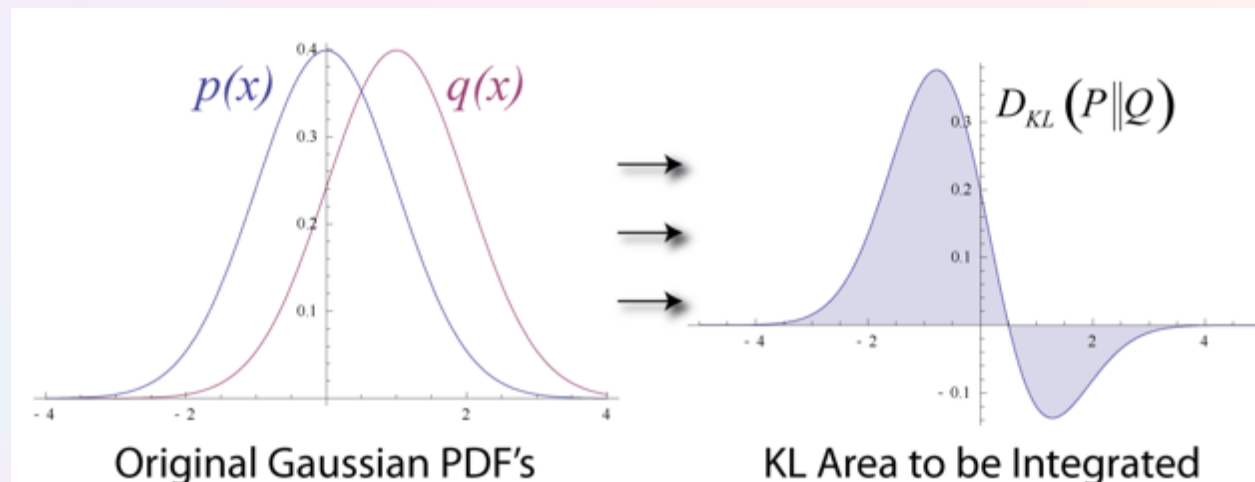
# Informally, **Kullback-Leibler Divergence (KL)** Measures the Distance Between Distributions

- Formally, **KL divergence** for discrete variables:

$$KL(P(\cdot), Q(\cdot)) = \mathbb{E}_{x \sim P} \left[ \log \frac{P(x)}{Q(x)} \right] = \sum_x P(x) \log \frac{P(x)}{Q(x)}$$

- Formally, **KL divergence** for continuous variables:

$$KL(p(\cdot), q(\cdot)) = \mathbb{E}_{X \sim p} \left[ \log \frac{p(x)}{q(x)} \right] = \int_x p(x) \log \frac{p(x)}{q(x)} dx$$



# Informally, **Kullback-Leibler Divergence (KL)** Measures the Distance Between Distributions

- $KL(p(\cdot), q(\cdot)) = \mathbb{E}_{X \sim p} \left[ \log \frac{p(x)}{q(x)} \right] = \int_x p(x) \log \frac{p(x)}{q(x)} dx$
- **Not symmetric!**
  - $KL(p(\cdot), q(\cdot)) \neq KL(q(\cdot), p(\cdot))$
- **Non-negative property**
  - $KL(p(\cdot), q(\cdot)) \geq 0$
- **Equal distribution property:**
  - $KL(p(\cdot), q(\cdot)) = 0 \iff p(\cdot) = q(\cdot)$

# One Use of KL Divergence Is to Estimate Distribution Parameters Only From Samples

- Let  $p(x)$  denote the **real/true** distribution of the data.
  - $p(x)$  is **unknown**.
- We only have samples  $\{x_i\}_{i=1}^n$  from  $p(x)$ .
- Let  $\hat{q}(x; \theta)$  denote an **estimate** of the true distribution.
  - Parametrized by  $\theta$ .
- We want to find  $\hat{q}(x; \theta)$  that is closest to  $p(x)$ :

$$\theta^* = \arg \min_{\theta} KL(p(\cdot), \hat{q}(\cdot; \theta))$$

# One Use of KL Divergence Is to Estimate Distribution Parameters Only From Samples

- We want to find  $\hat{q}(x; \theta)$  that is closest to  $p(x)$ :

$$\theta^* = \arg \min_{\theta} KL(p(\cdot), \hat{q}(\cdot; \theta))$$

- Wait, but we don't know  $p(x)$ , how do we do this?
- Two main ideas for simplification:
  1. Constants with respect to (w.r.t.)  $\theta$  can be ignored.
  2. Full expectation replaced by empirical expectation.

# Derivation of Minimum KL Divergence With Samples

$$\begin{aligned} & \arg \min_{\theta} KL(p(\cdot), \hat{q}(\cdot; \theta)) \\ &= \arg \min_{\theta} \mathbb{E}_{X \sim p} \left[ \log \frac{p(x)}{\hat{q}(x; \theta)} \right] \\ &= \arg \min_{\theta} \left( -\mathbb{E}_{X \sim p} [\log \hat{q}(x; \theta)] + \mathbb{E}_{X \sim p} [\log p(x)] \right) \\ &= \arg \min_{\theta} \left( -\mathbb{E}_{X \sim p} [\log \hat{q}(x; \theta)] + C \right) \\ &\approx \arg \min_{\theta} -\hat{\mathbb{E}}_{X \sim p} [\log \hat{q}(x; \theta)] \\ &= \arg \min_{\theta} -\frac{1}{n} \sum_{i=1}^n \log \hat{q}(x_i; \theta) \end{aligned}$$

# Maximum Likelihood Estimation (MLE) Is Another Way to Estimate Distribution Parameters From Samples

- **Likelihood function**  $\mathcal{L}(\theta; \mathcal{D})$ : how likely (or probable) a dataset  $\mathcal{D} = \{x_i\}_{i=1}^n$  is under a distribution with parameters  $\theta$ .

$$\mathcal{L}(\theta; \mathcal{D}) = \hat{q}(x_1, x_2, \dots, x_n; \theta)$$

- If we assume samples (or observations) of dataset are **independent and identically distributed (iid)**, then:

$$\mathcal{L}(\theta; \mathcal{D}) = \prod_{i=1}^n \hat{q}(x_i; \theta)$$

- Often simplified to the **log-likelihood function**  $l(\theta; \mathcal{D})$ :

$$l(\theta; \mathcal{D}) = \log \mathcal{L}(\theta; \mathcal{D}) = \sum_{i=1}^n \log \hat{q}(x_i; \theta)$$

# Maximum Likelihood (MLE) Is Another Way to Estimate Distribution Parameters From Samples

- Optimize the following:

$$\theta^* = \arg \max_{\theta} l(\theta; \mathcal{D}) = \arg \max_{\theta} \sum_{i=1}^n \log \hat{q}(x_i; \theta)$$

- Equivalent to:

$$\theta^* = \arg \min_{\theta} -\frac{1}{n} \sum_{i=1}^n \log \hat{q}(x_i; \theta)$$

- Wait, doesn't that look familiar?
- **MLE equivalent to minimum KL divergence!**

# MLE of Multivariate Gaussian Can Be Computed Via Empirical Mean and Covariance Matrix

- The MLE estimate (or equivalently minimum KL divergence) is simply the empirical mean and covariance matrix:

$$\hat{\mu}_{MLE} = \frac{1}{n} \sum_{i=1}^n x_i$$

$$\hat{\Sigma}_{MLE} = \frac{1}{n} \sum_{i=1}^n (x_i - \hat{\mu}_{MLE})(x_i - \hat{\mu}_{MLE})^T$$

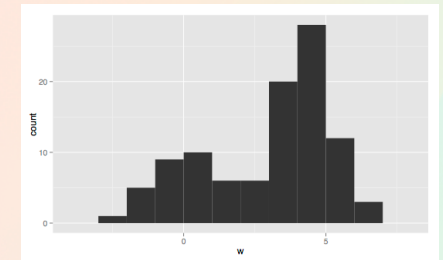
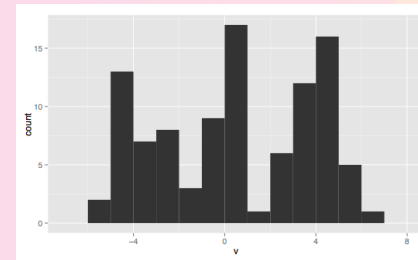
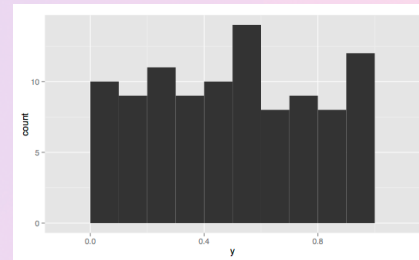
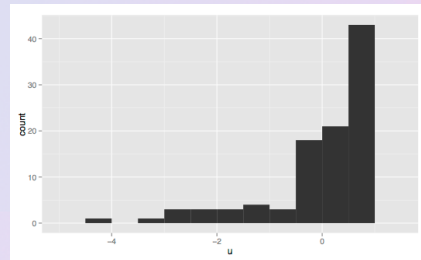
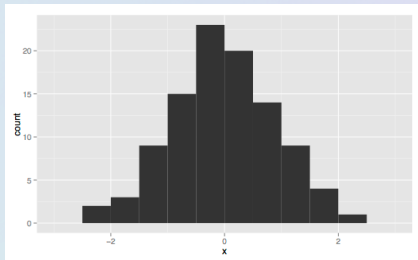
# Non-Parametric Density Estimation

# Non-Parametric Density Estimation

- Motivation
- Histograms
  - Choosing  $k$
  - Choosing bin edges
- Kernel density
  - Choosing bandwidth
  - Curse of dimensionality again

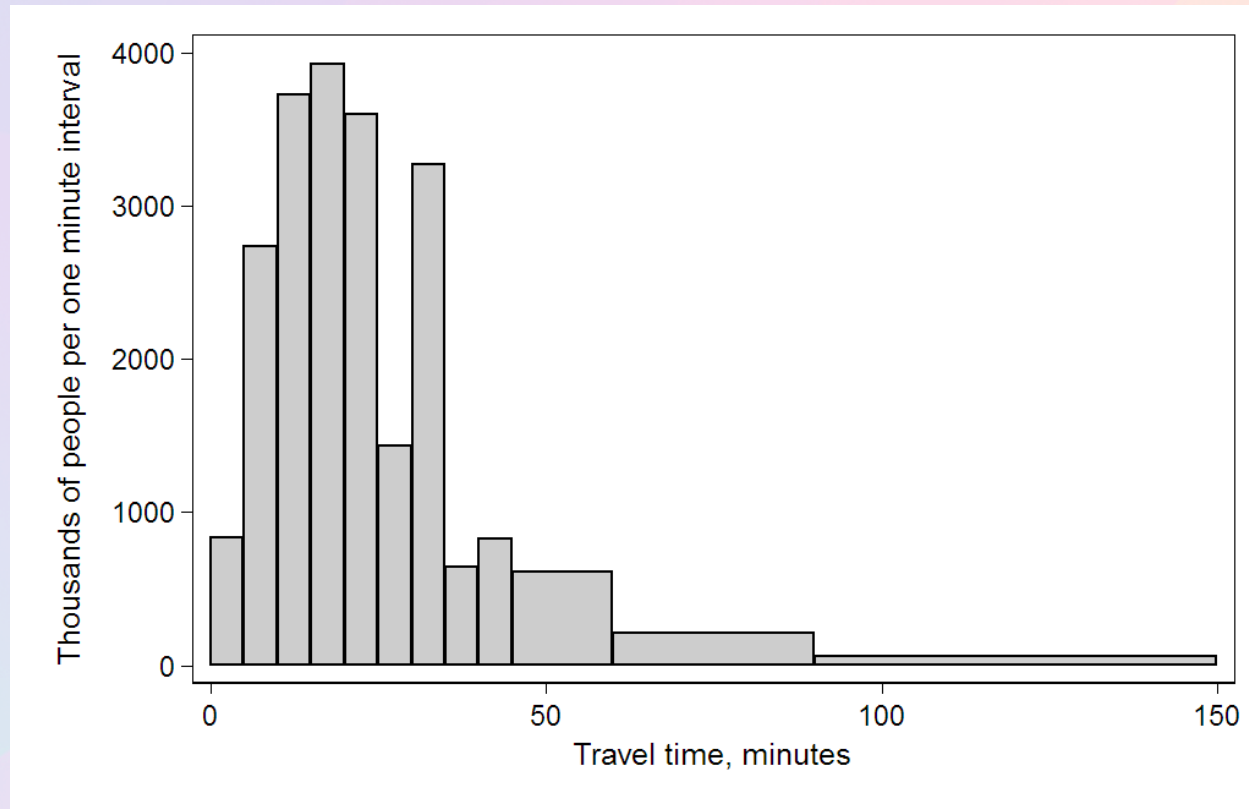
# Why Non-Parametric Density Estimates?

- Parametric densities are excellent if the assumptions are correct (e.g., Gaussian).
- However, the distributions may not align with the assumptions.

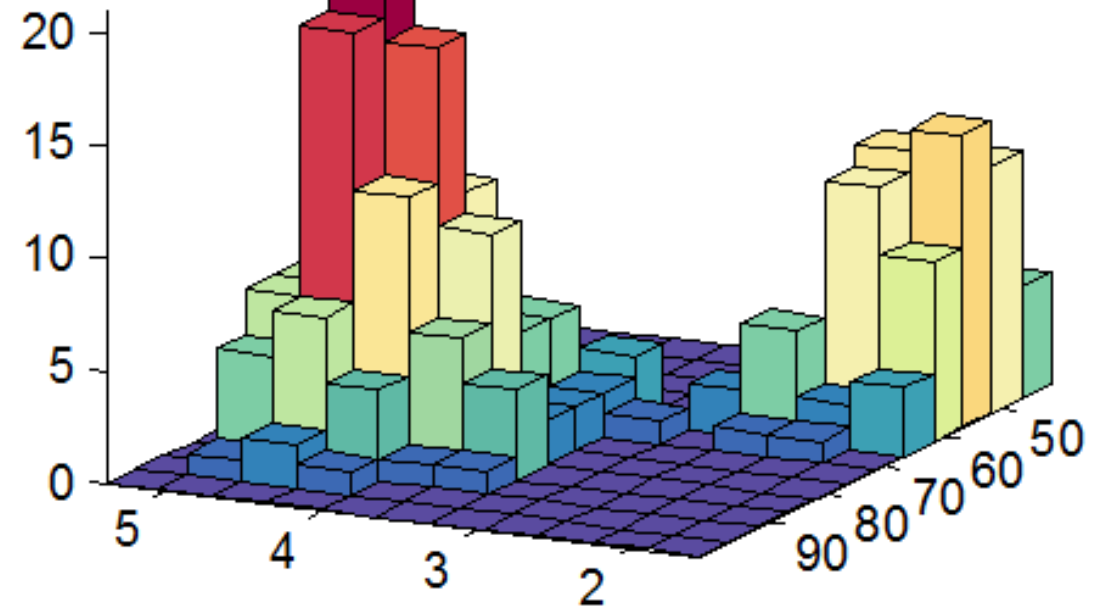
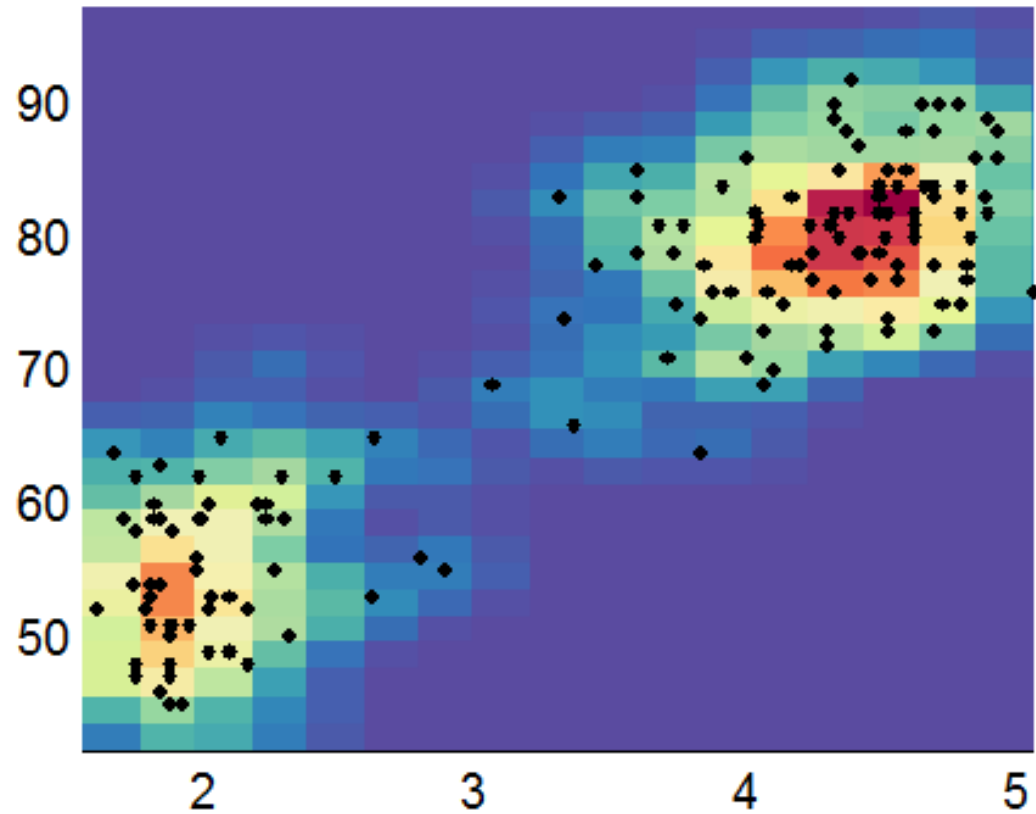


# Histograms Are the Simplest Density Estimators

- Setup bin locations
- Count number of samples that fall in each bin
- Normalize to be a density

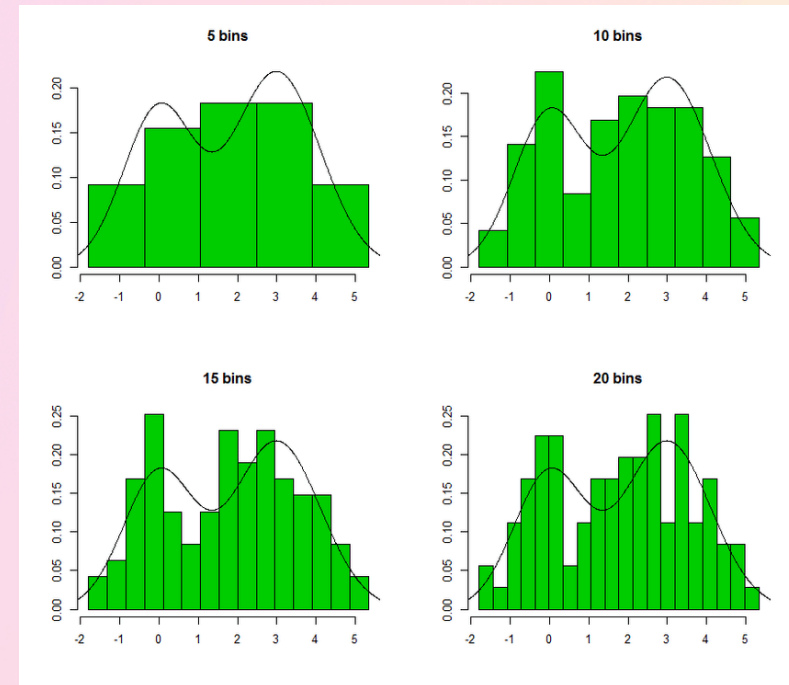


# 2d Histograms Can Be Created

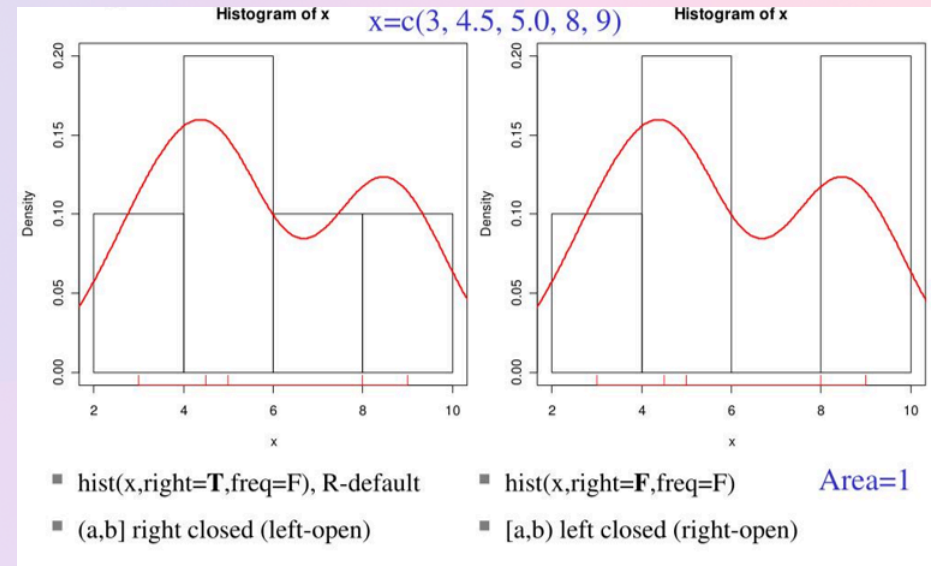


# How to Select the Number of Bins (Usually Denoted $K$ )?

- Too few bins will underfit.
- Too many bins will overfit.
- ML approach: **CV/Test** log likelihood.



# Drawbacks: Histograms Can Depend on Bin Edges and Are Not Smooth

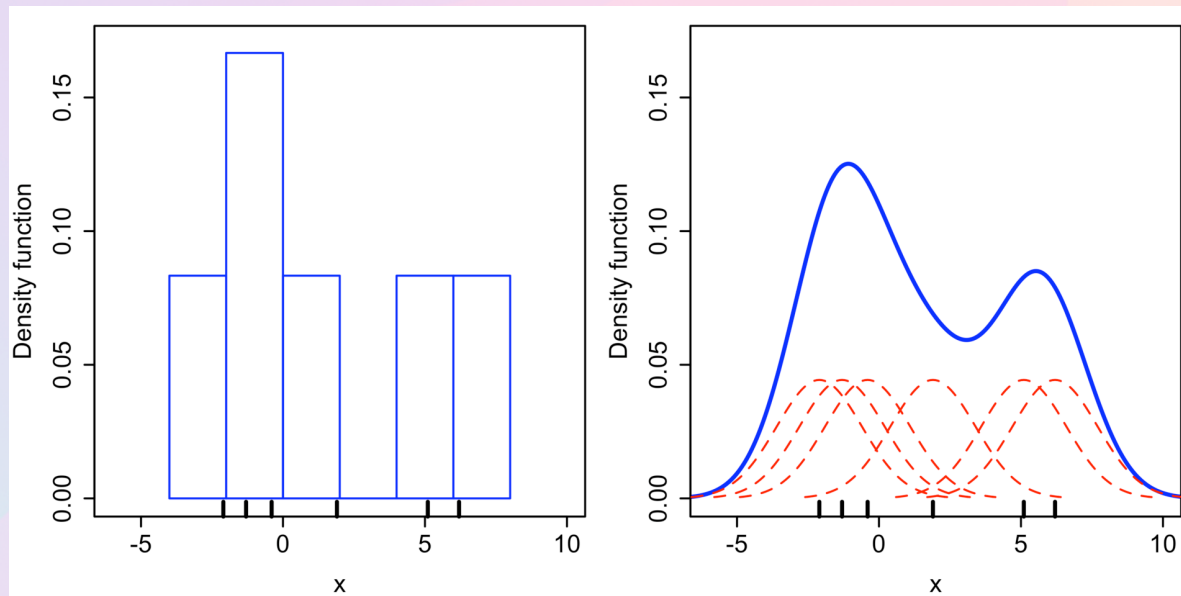


<https://www.slideserve.com/geona/introduction-to-non-parametric-statistics-kernel-density-estimation>

# Kernel Densities Overcome This Drawback by Placing a Gaussian Density at Each Point

- Kernel density has the following form:

$$\hat{p}(x) = \frac{1}{n} \sum_{i=1}^n p_{\text{base}}(x - x_i) = \frac{1}{n} \sum_{i=1}^n \mathcal{N}(x - x_i, \sigma^2)$$



# Similar to Number of Bins, the Key Parameter for Kernel Densities Is the “Bandwidth” or $\sigma$ Parameter

- Bandwidth can be selected via CV/Test log likelihood (similar to number of histogram bins).

