

# DSC 140A

*Probabilistic Modeling & Machine Learning*

Lecture 12 | Part 1

**Parametric Density Estimation**

# Bayes Classifier

- Recall the **Bayes Classifier**: predict

$$\begin{cases} 1, & \text{if } \mathbb{P}(Y = 1 \mid \vec{X} = \vec{x}) > \mathbb{P}(Y = 0 \mid \vec{X} = \vec{x}), \\ 0, & \text{otherwise.} \end{cases}$$

- Equivalently, using **Bayes' rule**:

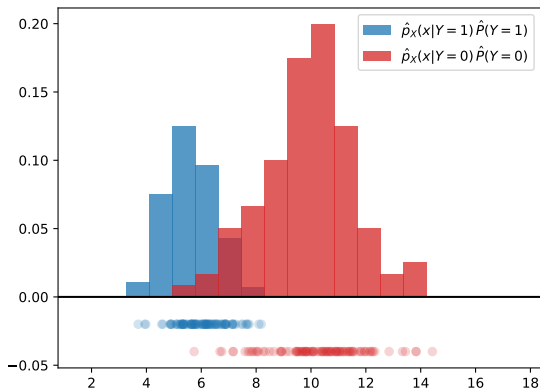
$$\begin{cases} 1, & \text{if } p_X(x \mid Y = 1)\mathbb{P}(Y = 1) > p_X(x \mid Y = 0)\mathbb{P}(Y = 0), \\ 0, & \text{otherwise.} \end{cases}$$

# Estimating Densities

- ▶ We rarely know the true distribution.
- ▶ We must **estimate** it from data.
- ▶ When  $\vec{X}$  is continuous, we estimate **density**.

# Last Time: Histogram Estimators

- **Histograms** provide one way of estimating densities.



# Histogram Drawbacks

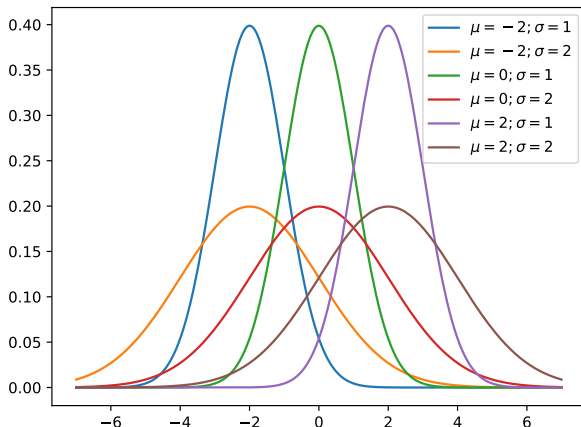
- ▶ We saw that histograms need massive amounts of data in high dimensions.
- ▶ The **Curse of Dimensionality**.

# Observation

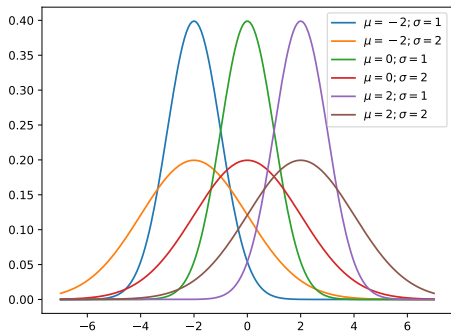
- ▶ Histogram estimators assume nothing about the **shape** of the true density.
- ▶ This makes them very flexible, but also data-hungry.
- ▶ **Idea:** Assume that the true, underlying density has a certain form.

# Example: Gaussians

- Often assume that the true distribution is **Gaussian** (aka, **Normal**).



# Example: Gaussians



- **Recall:** the pdf of the Gaussian distribution:

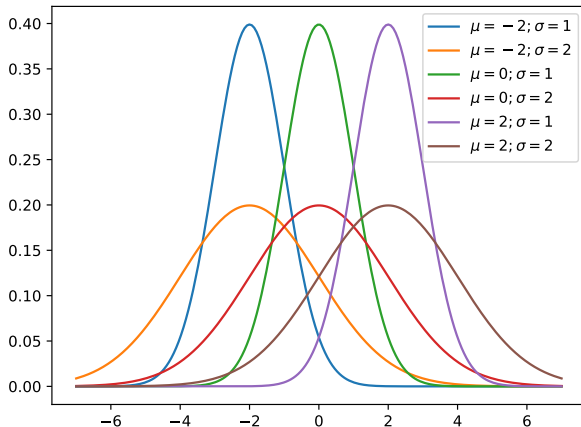
$$p(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

- $\mu$  and  $\sigma$  are **parameters**
  - $\mu$  controls center
  - $\sigma$  controls width



# Gaussian

- ▶ **Central Limit Theorem:** sums of independent random variables are Gaussian
- ▶ **Examples:** test scores, heights, measurement errors, ...



# Parametric Distributions

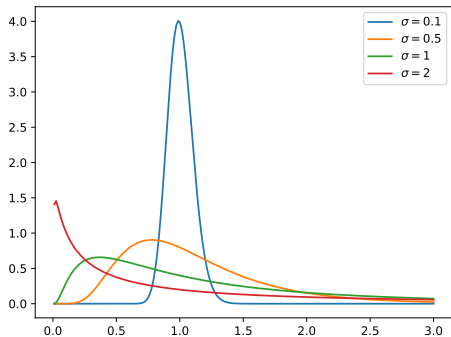
- ▶ A **parametric distribution** is **totally** determined by a finite number of **parameters**.
- ▶ **Example:** knowing  $\mu$  and  $\sigma$  tells you everything about a Gaussian distribution.

# Other Parametric Distributions

- ▶ There are many parametric distributions.
- ▶ **Discrete:** Bernoulli, Multinomial, Poisson, ...
- ▶ **Continuous:** Log-normal, Gamma, Pareto, ...

# Example: Lognormal

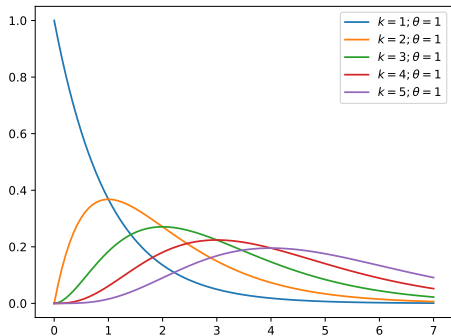
- ▶ Product of many independent positive random numbers.
- ▶ **Example:** length of comments in an internet forum



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

# Example: Gamma

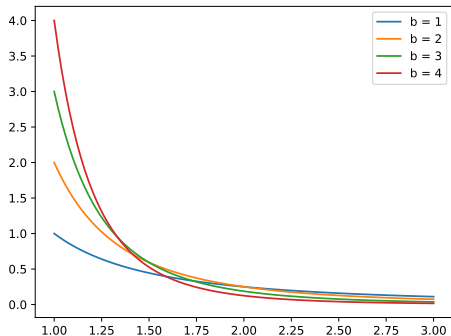
- **Examples:** wait times, size of rainfalls, insurance claims, ...



$$p(x; k, \theta) = \frac{1}{\Gamma(k)\theta^k} x^{k-1} e^{-x/\theta}$$

# Example: Pareto

- **Examples:** distribution of wealth, size of meteorites, ...



$$p(x; x_m, \alpha) = \frac{\alpha x_m^\alpha}{x^{\alpha+1}}$$

# Parametric Density Estimation

- ▶ In **parametric density estimation**, we assume data comes from some parametric density.
  - ▶ E.g., Gaussian, Log-Normal, Pareto, etc.
- ▶ But we don't know the parameters.
- ▶ Use data to **estimate** the parameters.

# Non-Parametric Density Estimation

- ▶ Contrast this with estimating density with histograms.
- ▶ There were no parameters controlling the shape of the density.
- ▶ Histograms are **non-parametric** density estimators.



# DSC 140A

*Probabilistic Modeling & Machine Learning*

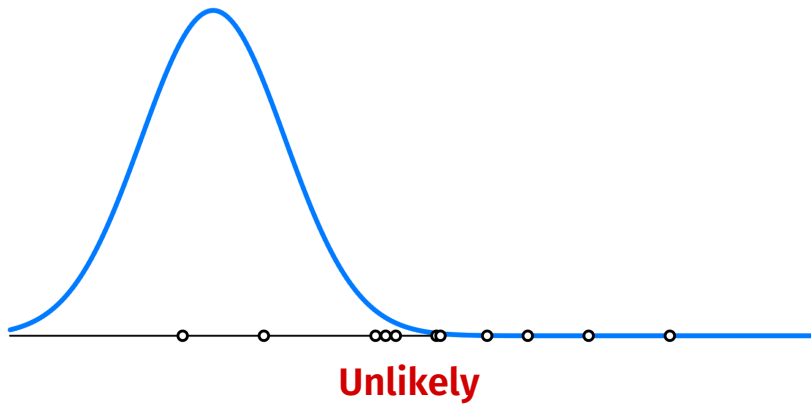
Lecture 12 | Part 2

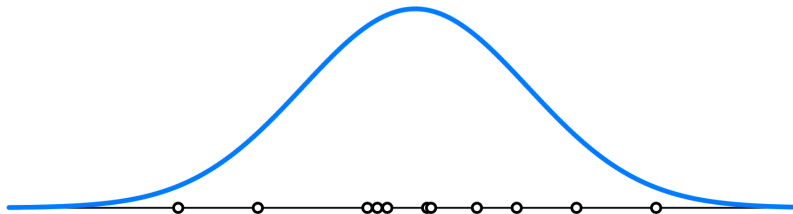
**Maximum Likelihood Estimation**

# Parametric Density Estimation

- ▶ Suppose we have data  $x^{(1)}, \dots, x^{(n)} \in \mathbb{R}$ .
- ▶ Assume it came from a parametric distribution.
  - ▶ Say, a Gaussian.
- ▶ What were the parameter values used to generate the data?
- ▶ Using data to guess  $\mu$  and  $\sigma$  is called **estimating** the parameters.







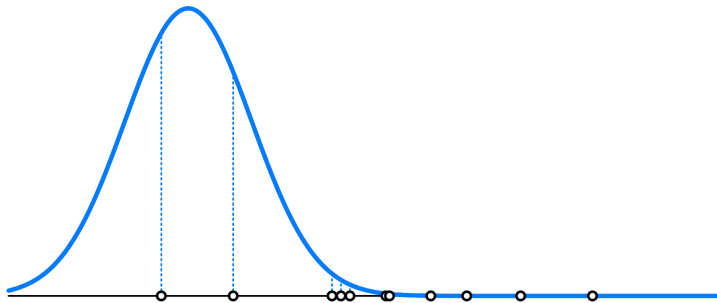
Likely

# Intuition

- ▶ Some parameter choices seem **more likely** than others.
- ▶ That is, there is a greater chance that the data could have been generated by them.
- ▶ How can we quantify this?

# Intuition

- ▶ Let  $p$  be the Gaussian probability density function.
- ▶  $p(x^{(i)}; \mu, \sigma)$  quantifies how likely it is to see  $x^{(i)}$  if parameters  $\mu$  and  $\sigma$  are used.



## Exercise

Assume that  $x^{(1)}, \dots, x^{(n)}$  are all sampled independently from a density with parameters  $\mu, \sigma$ .

Think of  $p(x^{(i)}; \mu, \sigma)$  as the “chance” of seeing  $x^{(i)}$  under parameters  $\mu$  and  $\sigma$ .

What is the chance of seeing  $x^{(1)}$  and  $x^{(2)}$  and  $x^{(3)}$  and ... and  $x^{(n)}$ ?



# Intuition

- ▶  $p(x^{(1)}; \mu, \sigma) \times p(x^{(2)}; \mu, \sigma) \times \dots \times p(x^{(n)}; \mu, \sigma)$  quantifies likelihood of seeing  $x^{(1)}, \dots, x^{(n)}$  simultaneously.
- ▶ In fact, it is the **joint density** of the data.
- ▶ But instead think of this as a function of  $\mu$  and  $\sigma$ .

# Likelihood

- The **likelihood** of  $\mu$  and  $\sigma$  with respect to data  $x^{(1)}, \dots, x^{(n)}$  is:

$$\begin{aligned}\mathcal{L}(\mu, \sigma; x^{(1)}, \dots, x^{(n)}) &= p(x^{(1)}; \mu, \sigma) \times p(x^{(2)}; \mu, \sigma) \times \dots \times p(x^{(n)}; \mu, \sigma) \\ &= \prod_{i=1}^n p(x^{(i)}; \mu, \sigma)\end{aligned}$$

# Likelihood

- ▶ The likelihood function takes in parameters  $\mu$  and  $\sigma$  and returns a real number.
- ▶ **Interpretation:** likelihood that data was generated by this choice of  $\mu$  and  $\sigma$ .
- ▶ **Goal:** find  $\mu$  and  $\sigma$  that **maximize** the likelihood.

<http://dsc140a.com/static/vis/mle/>

# Maximizing Likelihood

- ▶ To maximize  $\mathcal{L}(\mu, \sigma)$ , we might take derivatives  $\frac{\partial \mathcal{L}}{\partial \mu}$  and  $\frac{\partial \mathcal{L}}{\partial \sigma}$ , set to 0, solve.
- ▶ But the likelihood is often difficult to work with.

# Example: Gaussian

- Assume that  $p$  is the Gaussian pdf.

$$p(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

- Then the likelihood function is:

$$\mathcal{L}(\mu, \sigma) = \prod_{i=1}^n \left( \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x^{(i)}-\mu)^2}{2\sigma^2}} \right)$$

# Log Likelihood

- It is typically easier to work with the **log likelihood** instead.

$$\tilde{\mathcal{L}}(\mu, \sigma) = \ln \mathcal{L}(\mu, \sigma)$$

- **Fact:** Because  $\ln x$  is **monotonically increasing**, a maximizer of  $\ln \mathcal{L}$  also maximizes  $\mathcal{L}$

## Procedure: Gaussian

1. Write the log likelihood function  $\tilde{\mathcal{L}}$ .
2. Take derivatives  $\partial\tilde{\mathcal{L}}/\partial\mu$  and  $\partial\tilde{\mathcal{L}}/\partial\sigma$
3. Set to zero and solve for  $\mu$  and  $\sigma$ .



## Recall: Log Properties

- ▶ If  $a$  and  $b$  are positive:  $\ln(a \times b) = \ln a + \ln b$
- ▶ If  $a$  and  $b$  are positive:  $\ln(a/b) = \ln a - \ln b$
- ▶ If  $a$  is positive:  $\ln a^p = p \ln a$

# Step 1: Write Log Likelihood

- ▶ Write the log likelihood function for the Normal distribution.

## Step 2: Differentiate

- ▶ We have:  $\tilde{\mathcal{L}} = \sum_{i=1}^n \left[ -\ln \sigma - \ln \sqrt{2\pi} - \frac{(x^{(i)} - \mu)^2}{2\sigma^2} \right]$
- ▶ Compute  $\partial \tilde{\mathcal{L}} / \partial \mu$ :

## Step 2: Differentiate

- ▶ We have:  $\tilde{\mathcal{L}} = \sum_{i=1}^n \left[ -\ln \sigma - \ln \sqrt{2\pi} - \frac{(x^{(i)} - \mu)^2}{2\sigma^2} \right]$
- ▶ Compute  $\partial \tilde{\mathcal{L}} / \partial \sigma$ :

## Step 3: Solve

- ▶ We have  $\partial \tilde{L} / \partial \mu = \frac{1}{\sigma^2} \sum_{i=1}^n (x^{(i)} - \mu)$
- ▶ Solve  $\partial \tilde{L} / \partial \mu = 0$  for  $\mu$ .

## Step 3: Solve

- ▶ We have  $\partial \tilde{L} / \partial \sigma = \sum_{i=1}^n \left[ -\frac{1}{\sigma} + \frac{(x^{(i)} - \mu)^2}{\sigma^3} \right]$
- ▶ Solve  $\partial \tilde{L} / \partial \sigma = 0$  for  $\sigma$ .

# MLEs for Gaussian Distribution

- We have found the **maximum likelihood estimates** for the Gaussian distribution:

$$\mu_{\text{MLE}} = \frac{1}{n} \sum_{i=1}^n x^{(i)} \quad \sigma_{\text{MLE}} = \sqrt{\frac{1}{n} \sum_{i=1}^n (x^{(i)} - \mu_{\text{MLE}})^2}$$

# “Fitting” a Gaussian

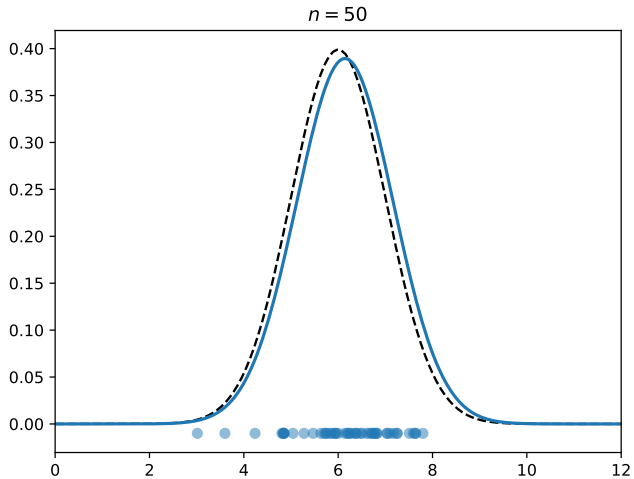
- ▶ Suppose we wish to “fit” a Gaussian to data  $x^{(1)}, \dots, x^{(n)}$ .
- ▶ The **maximum likelihood** approach:
  1. Compute:

$$\mu_{\text{MLE}} = \frac{1}{n} \sum_{i=1}^n x^{(i)} \quad \sigma_{\text{MLE}} = \sqrt{\frac{1}{n} \sum_{i=1}^n (x^{(i)} - \mu_{\text{MLE}})^2}$$

2. Use these as parameters of the Gaussian.



# Example



# In General

- ▶ **Maximum Likelihood Estimation** (MLE) can be used for a variety of densities.
- ▶ Suppose density  $p$  has parameters  $\theta_1, \dots, \theta_k$

1. Write log likelihood function:

$$\ln \mathcal{L}(\theta_1, \dots, \theta_k) = \sum_{i=1}^n \ln p(x^{(1)}, \dots, x^{(n)}; \theta_1, \dots, \theta_k)$$

2. Compute derivatives:  $\partial \tilde{\mathcal{L}} / \partial \theta_1, \partial \tilde{\mathcal{L}} / \partial \theta_2, \dots, \partial \tilde{\mathcal{L}} / \partial \theta_k$
3. Set derivatives to zero, solve for  $\theta_1, \dots, \theta_k$ .

## In Practice

- ▶ The MLE for a parameter only needs to be derived once.
- ▶ Many textbooks, statistics packages, and Wikipedia list the MLE parameter estimators.

# DSC 140A

*Probabilistic Modeling & Machine Learning*

Lecture 12 | Part 3

**Parametric vs. Non-Parametric Density Estimation**

# Making Predictions

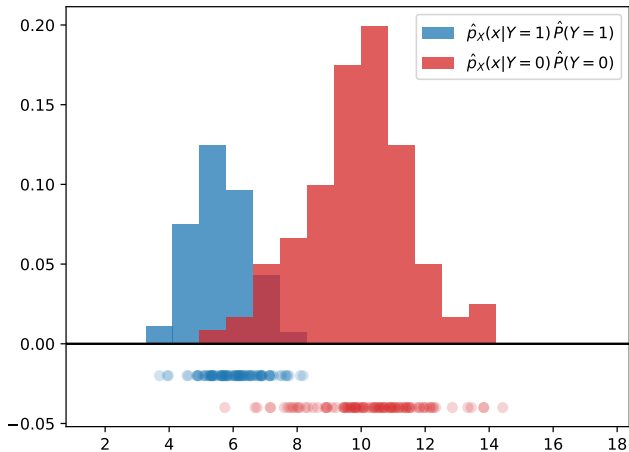
- ▶ We observe a data set  $\{(x^{(i)}, y_i)\}$  of flipper lengths and penguin species (0 or 1).
- ▶ **Task:** Given the flipper length of a new penguin, what is its species?
- ▶ Bayes' classifier: predict
$$\begin{cases} 1, & \text{if } p_X(x | Y = 1)\mathbb{P}(Y = 1) > p_X(x | Y = 0)\mathbb{P}(Y = 0), \\ 0, & \text{otherwise.} \end{cases}$$

# Estimating Densities

- ▶ We must estimate  $p_X(x | Y = 0)$  and  $p_X(x | Y = 1)$ .
- ▶ Approach 1: Non-parametric (histograms)
- ▶ Approach 2: Parametric

# Approach 1: Non-Parametric

- Estimate  $p_X(x | Y = 0)$  and  $p_X(x | Y = 1)$  with histograms.



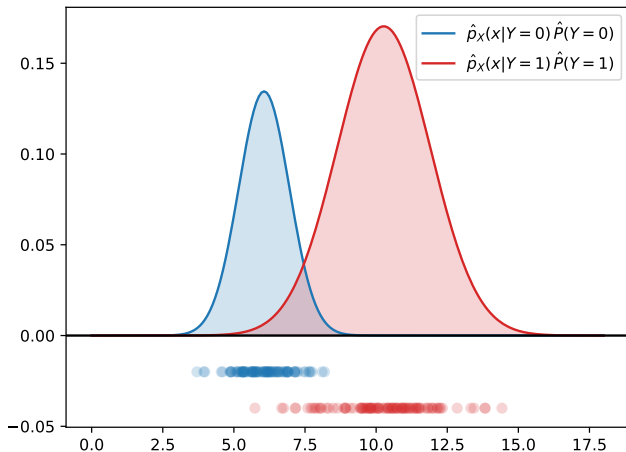
## Approach 2: Parametric

- ▶ Must choose a parametric distribution.
- ▶ Plotting a histogram, data looks roughly normal.
- ▶ We will fit Gaussians.



# Approach 2: Parametric

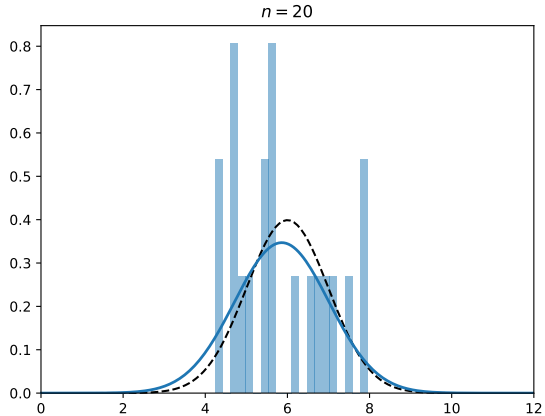
- Estimate  $p_X(x | Y = 0)$  and  $p_X(x | Y = 1)$  by fitting Gaussians with MLE.



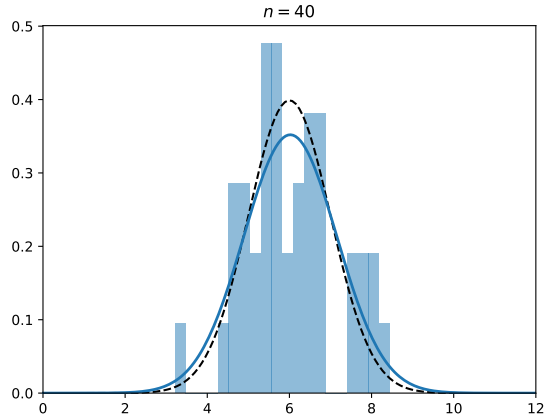
# Data Requirements

- ▶ Suppose the underlying distribution that produced the data actually was a Gaussian.
  - ▶ Or close to one.
- ▶ The parametric approach will require less data than the non-parametric.

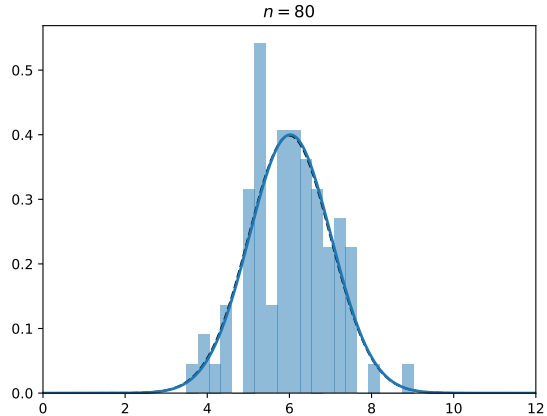
# Data Requirements



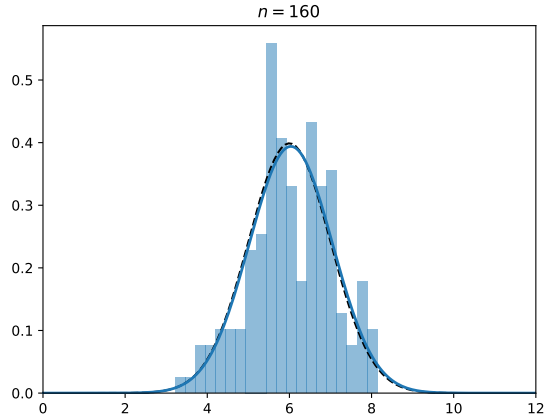
# Data Requirements



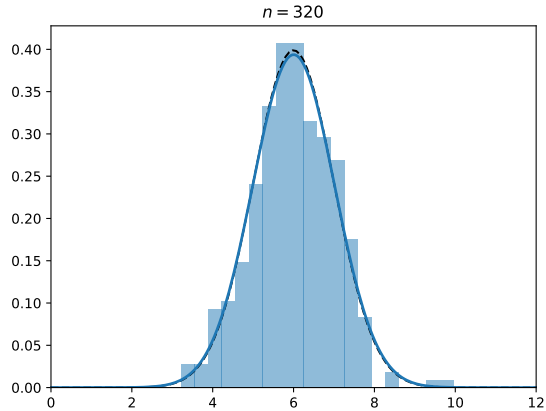
# Data Requirements



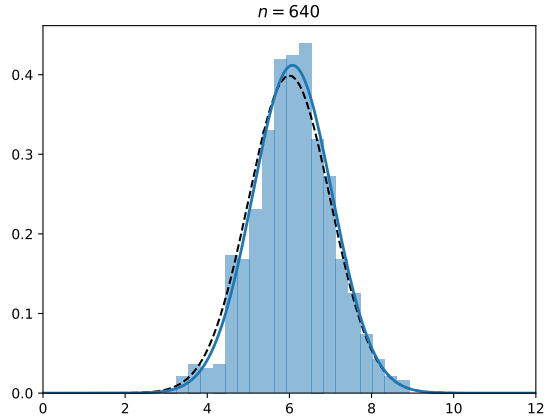
# Data Requirements



# Data Requirements

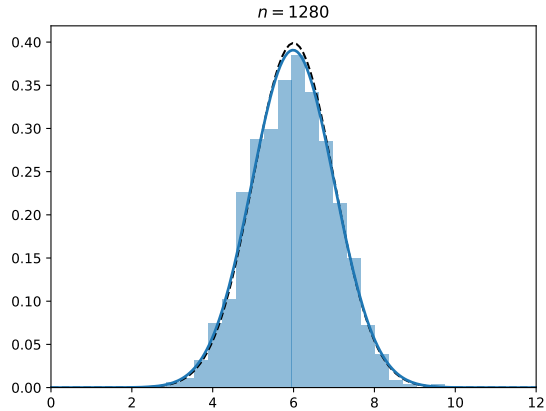


# Data Requirements





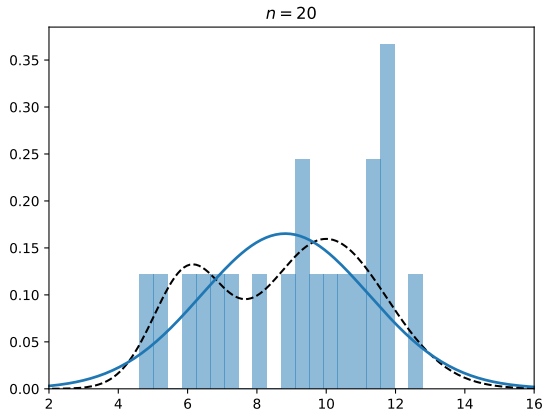
# Data Requirements



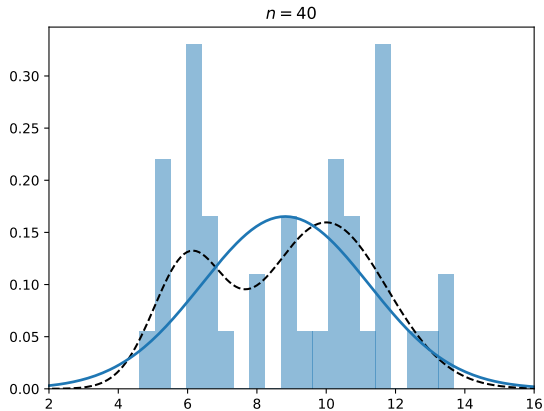
# Mis-specification

- ▶ However, suppose the underlying distribution is **not** Gaussian.
- ▶ No amount of data will allow the parametric approach to get close.
  - ▶ The model has been **mis-specified**.
- ▶ But the non-parametric approach will be close, eventually.

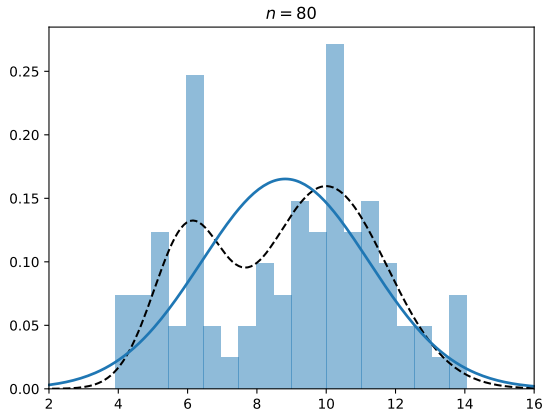
# Parametric vs. Non-Parametric



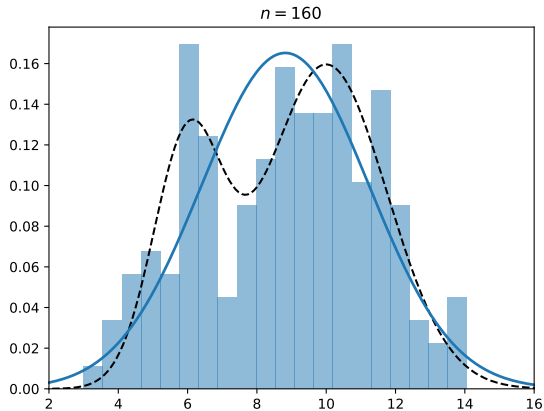
# Parametric vs. Non-Parametric



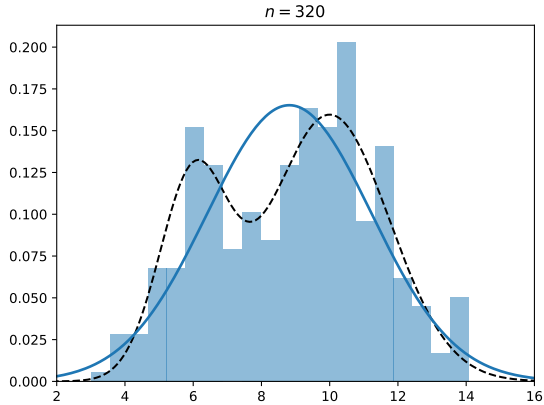
# Parametric vs. Non-Parametric



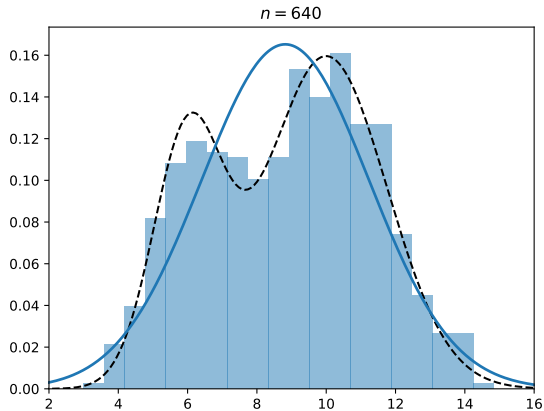
# Parametric vs. Non-Parametric



# Parametric vs. Non-Parametric

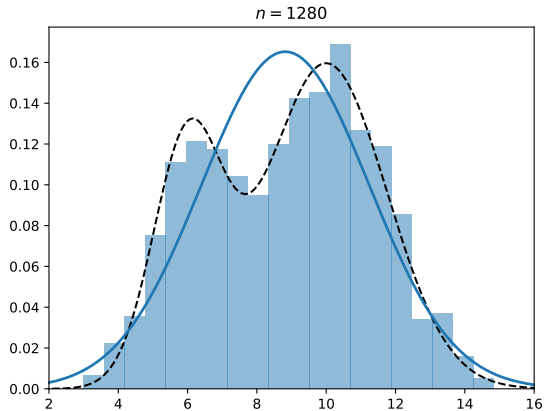


# Parametric vs. Non-Parametric





# Parametric vs. Non-Parametric



# High Dimensions

- ▶ Non-parametric approaches can fit arbitrary densities, but they require lots of data.
  - ▶ Especially in high dimensions!
- ▶ Parametric approaches require less data, provided that they are correctly specified.
- ▶ **Next time:** parametric density estimation in high dimensions.