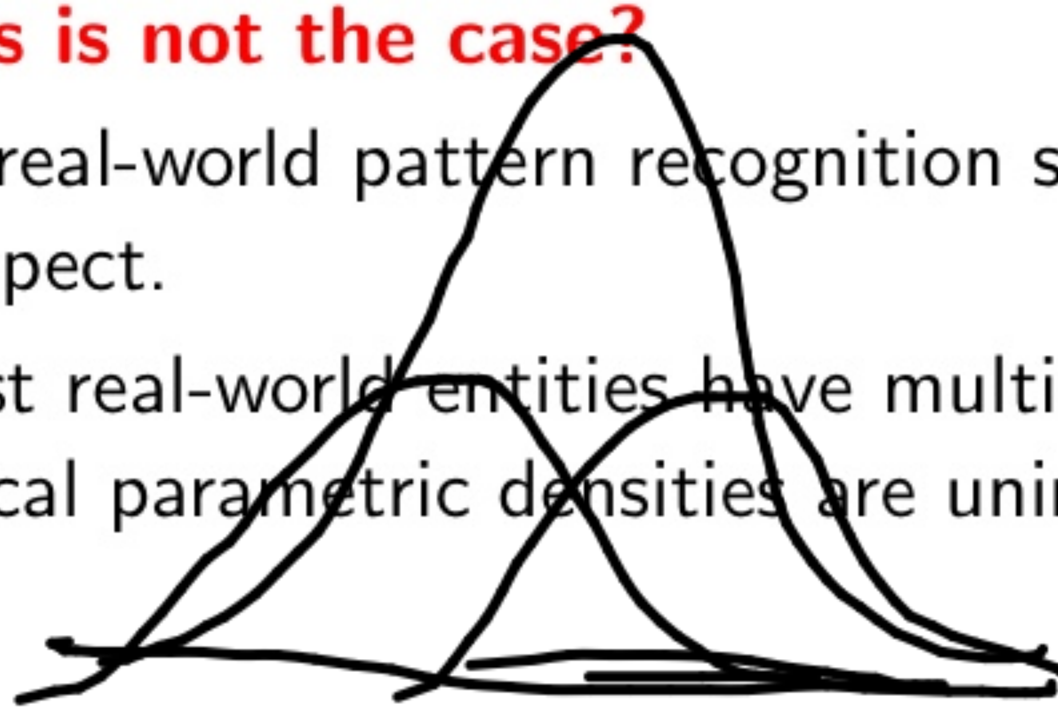
Nonparametric Methods

Jason Corso

SUNY at Buffalo

Nonparametric Methods Overview

- Previously, we've assumed that the forms of the underlying densities were of some particular known parametric form.
- **But, what if this is not the case?**
- Indeed, for most real-world pattern recognition scenarios this assumption is suspect.
- For example, most real-world entities have multimodal distributions whereas all classical parametric densities are unimodal.

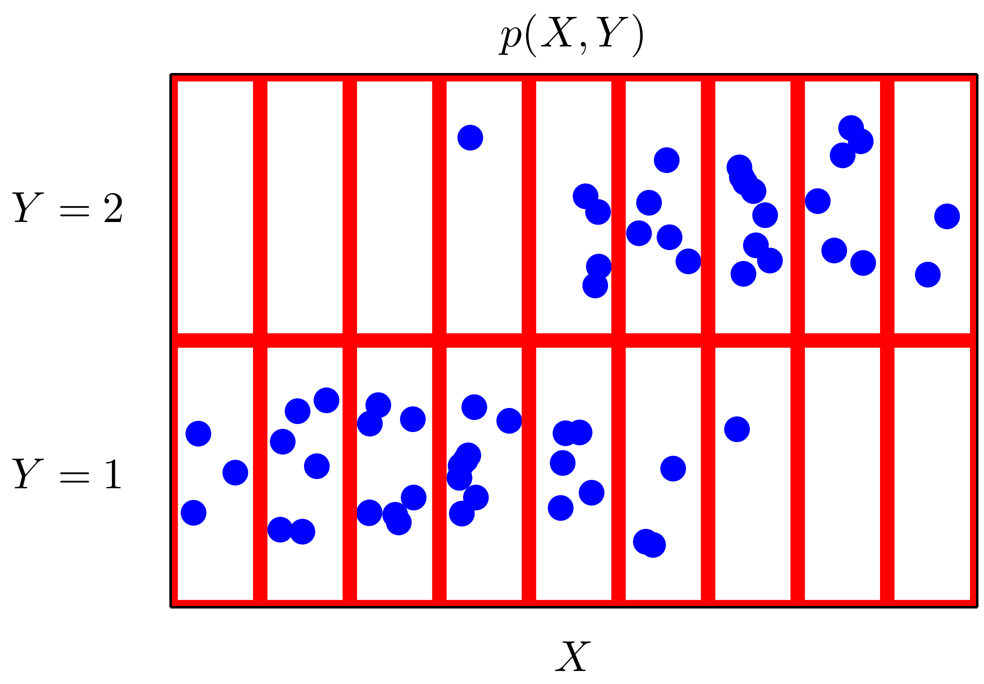


P —— MD —— NP

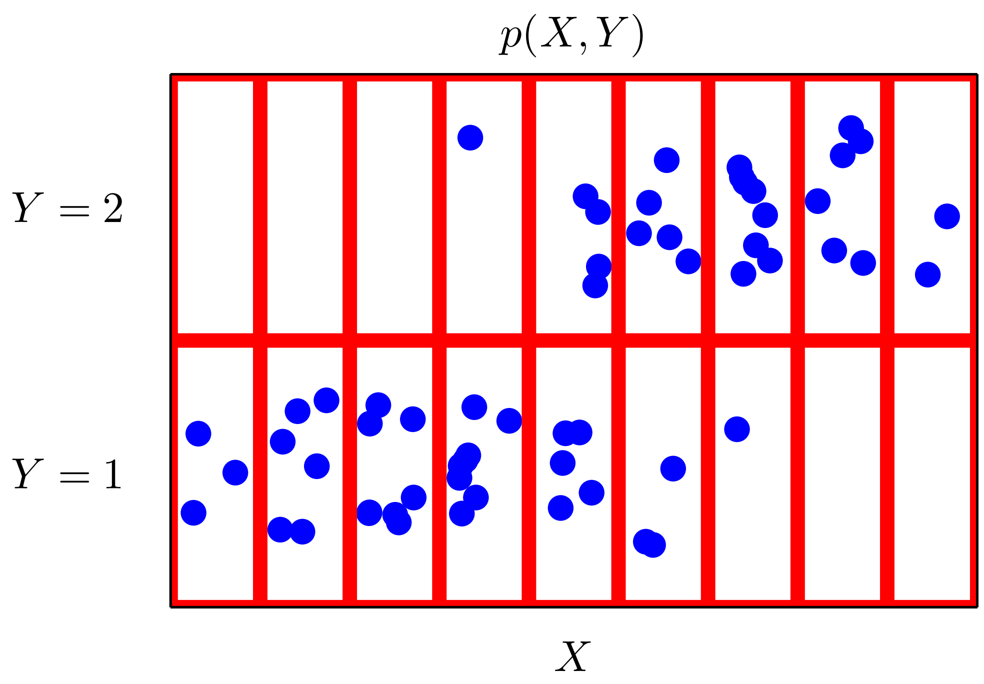
Nonparametric Methods Overview

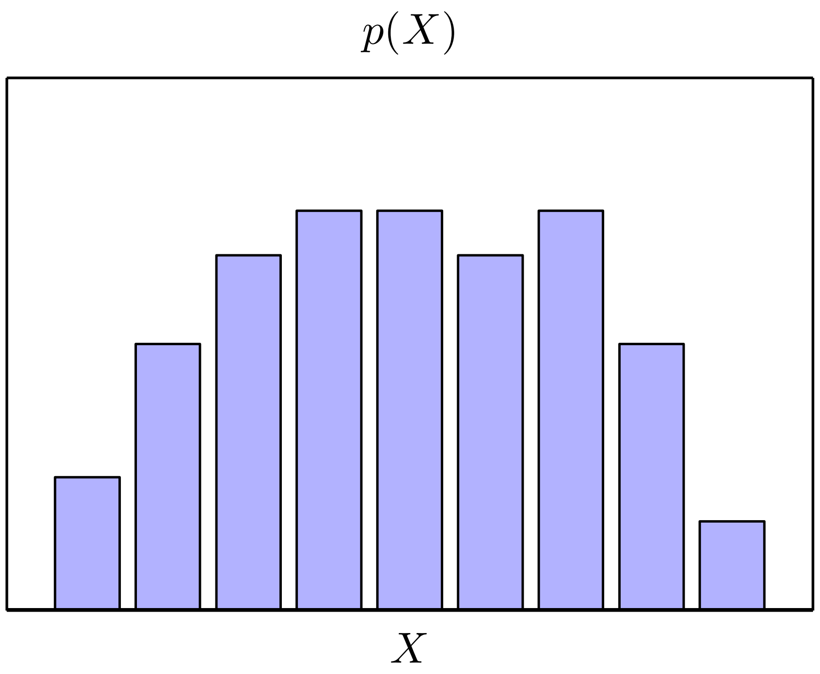
- Previously, we've assumed that the forms of the underlying densities were of some particular known parametric form.
- **But, what if this is not the case?**
- Indeed, for most real-world pattern recognition scenarios this assumption is suspect.
- For example, most real-world entities have multimodal distributions whereas all classical parametric densities are unimodal.
- We will examine **nonparametric** procedures that can be used with arbitrary distributions and without the assumption that the underlying form of the densities are known.
  - Histograms.
  - Kernel Density Estimation / Parzen Windows.
  - k-Nearest Neighbor Density Estimation.
  - Real Example in Figure-Ground Segmentation

Histograms

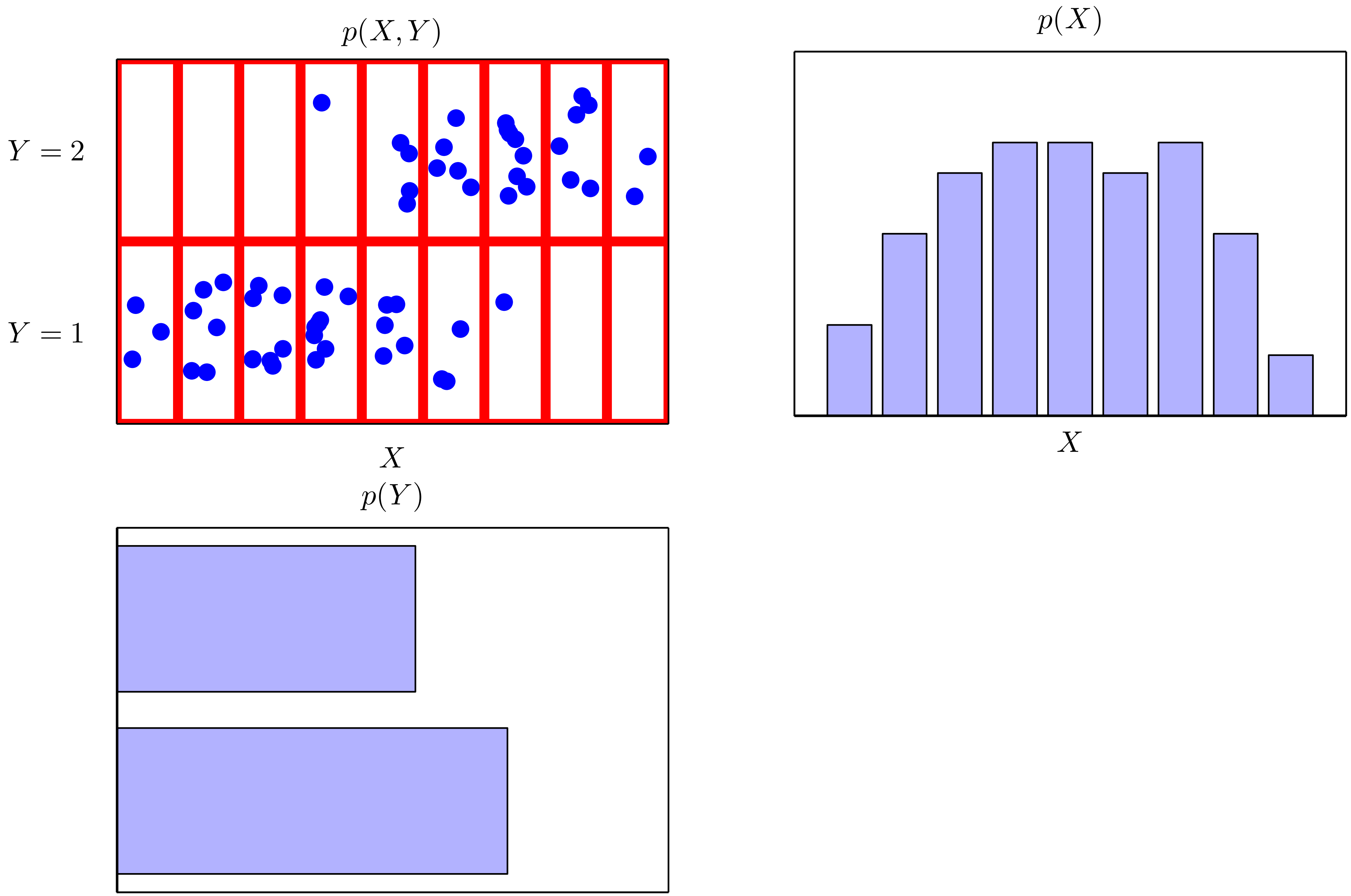


Histograms

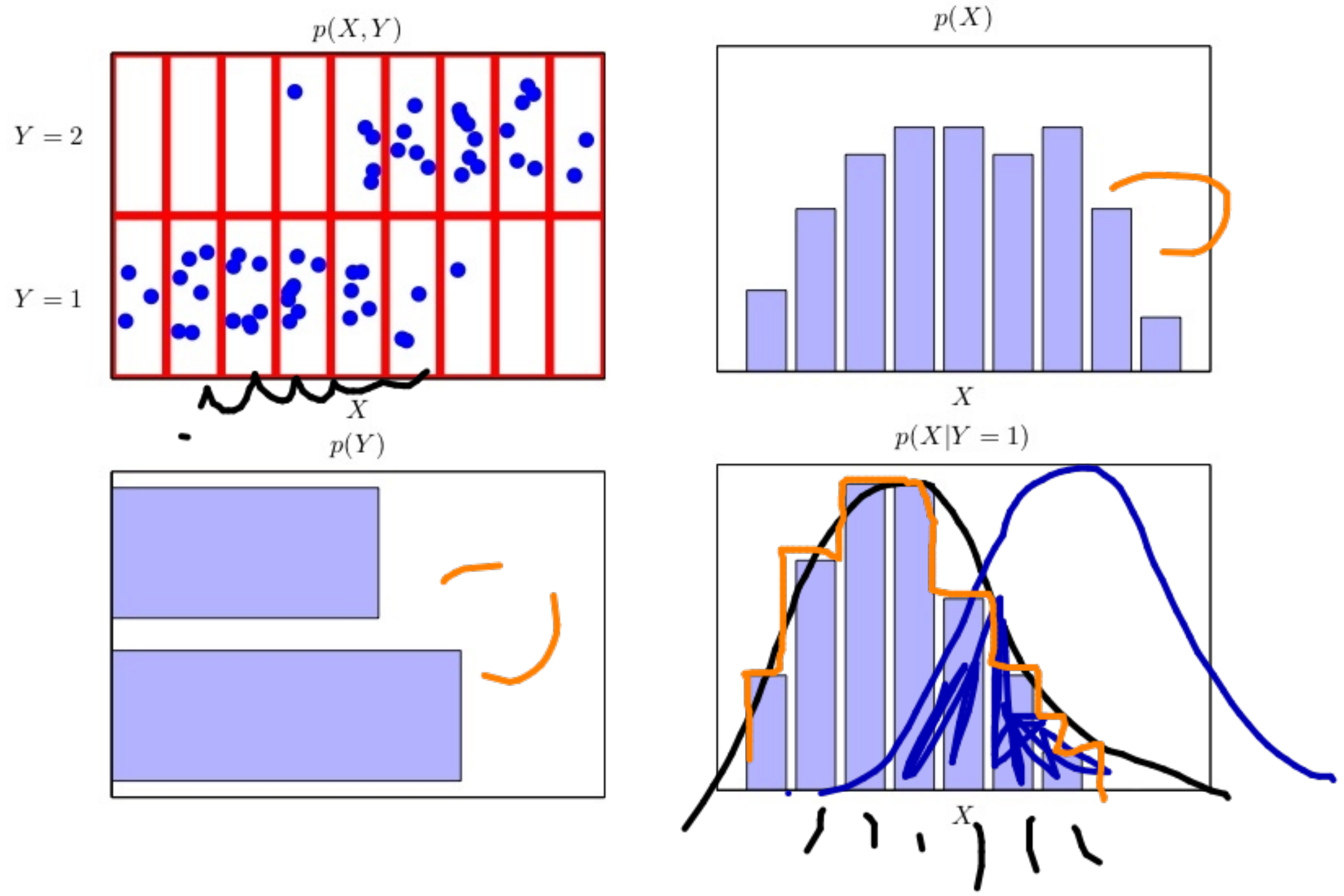




Histograms



Histograms



Histogram Density Representation

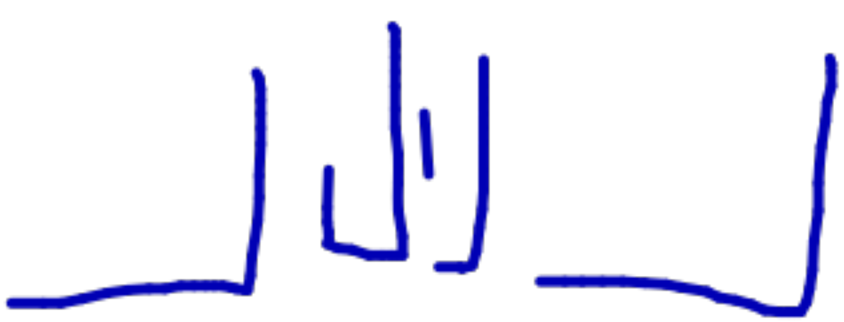
- Consider a single continuous variable x and let's say we have a set $\mathcal{D}$ of N of them $\{x_1, \ldots, x_N\}$. Our goal is to model $p(x)$ from $\mathcal{D}$.

Histogram Density Representation

- Consider a single continuous variable x and let's say we have a set $\mathcal{D}$ of N of them $\{x_1, \ldots, x_N\}$. Our goal is to model $p(x)$ from $\mathcal{D}$.
- Standard histograms simply partition x into distinct bins of width Δ_i and then count the number n_i of observations x falling into bin i.

Histogram Density Representation

- Consider a single continuous variable x and let's say we have a set $\mathcal{D}$ of N of them $\{x_1, \ldots, x_N\}$. Our goal is to model $p(x)$ from $\mathcal{D}$.
- Standard histograms simply partition x into distinct bins of width Δ_i and then count the number n_i of observations x falling into bin i.
- To turn this count into a normalized probability density, we simply divide by the total number of observations N and by the width Δ_i of the bins.
- This gives us:



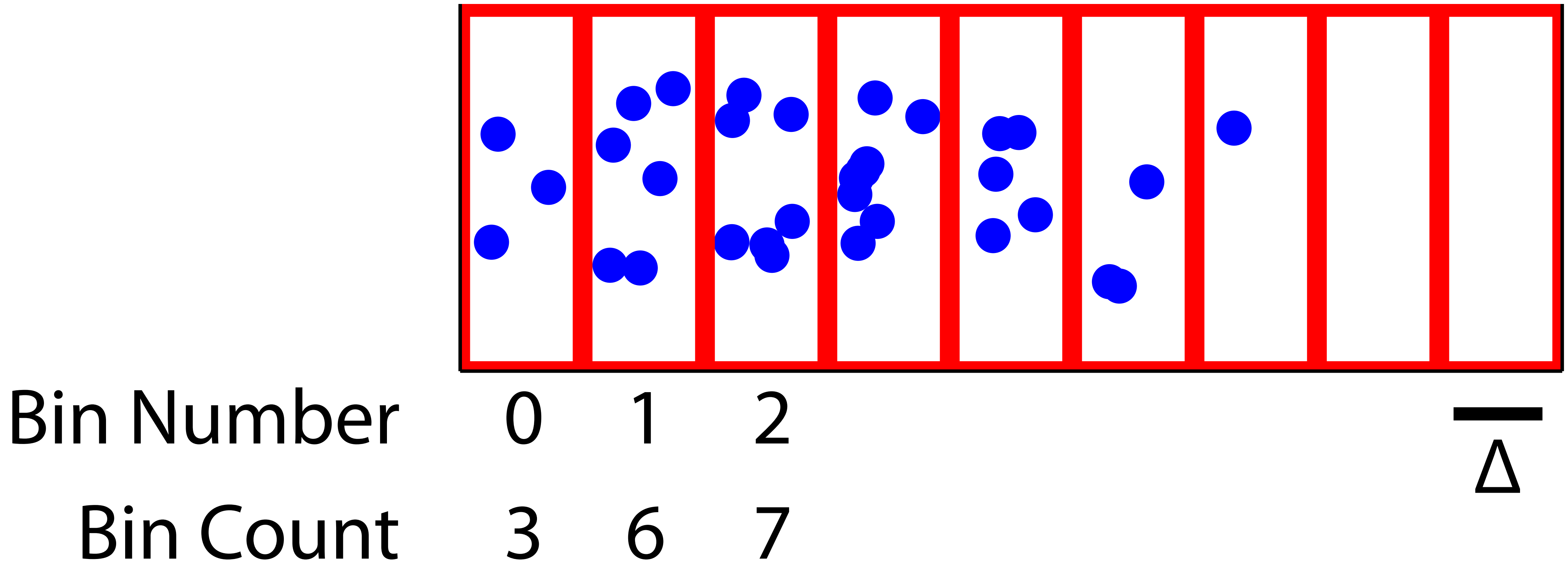
$$p_i = \frac{n_i}{N\Delta_i} \tag{1}$$

Histogram Density Representation

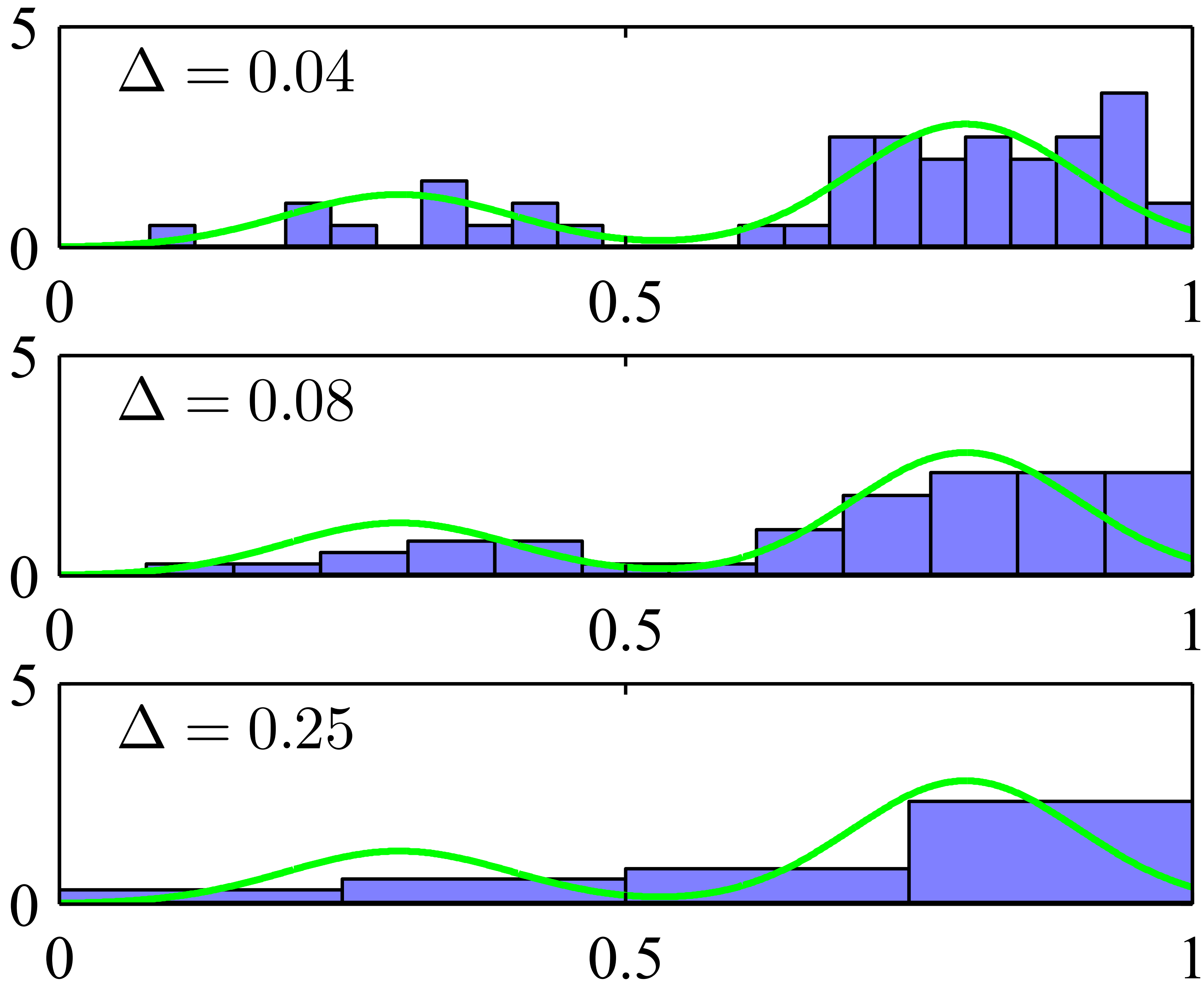
- Consider a single continuous variable x and let's say we have a set $\mathcal{D}$ of N of them $\{x_1, \dots, x_N\}$. Our goal is to model $p(x)$ from $\mathcal{D}$.
- Standard histograms simply partition x into distinct bins of width Δ_i and then count the number n_i of observations x falling into bin i.
- To turn this count into a normalized probability density, we simply divide by the total number of observations N and by the width Δ_i of the bins.
- This gives us:

$$p_i = \frac{n_i}{N\Delta_i} \tag{1}$$

- Hence the model for the density $p(x)$ is constant over the width of each bin. (And often the bins are chosen to have the same width $\Delta_i = \Delta$.)

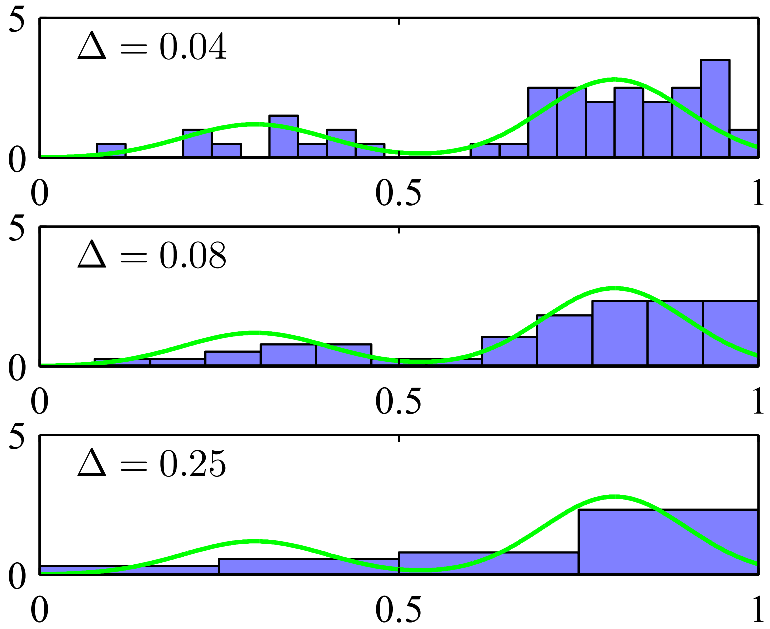

Bin Number 0 1 2
Bin Count 3 6 7
Δ

Histogram Density as a Function of Bin Width



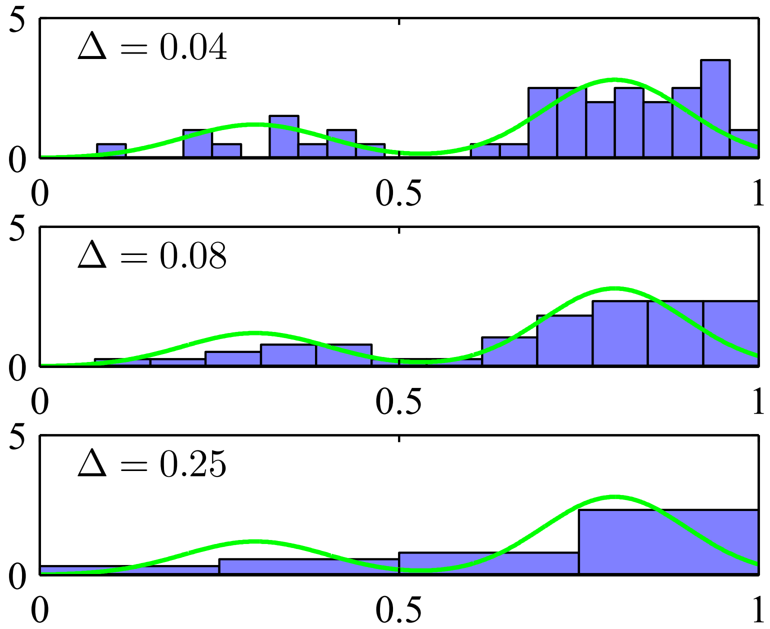
Histogram Density as a Function of Bin Width

- The green curve is the underlying true density from which the samples were drawn. It is a mixture of two Gaussians.



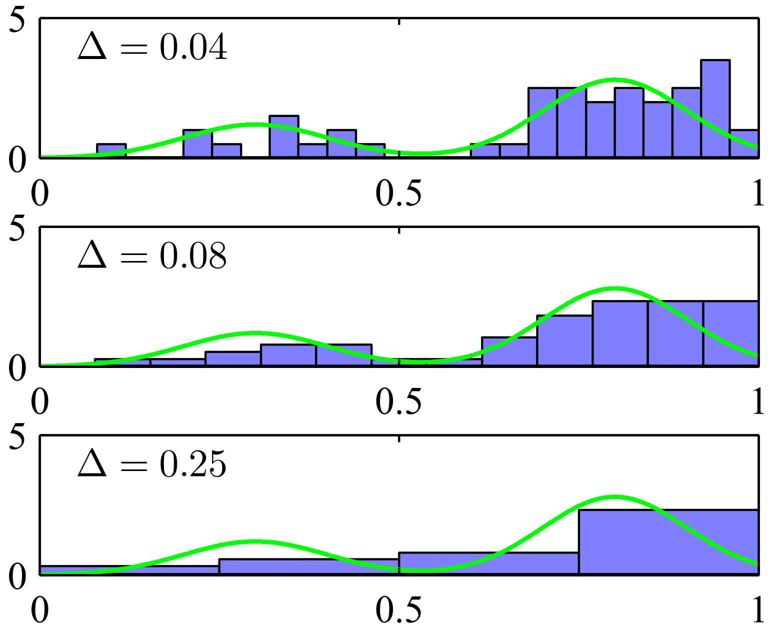
Histogram Density as a Function of Bin Width

- The green curve is the underlying true density from which the samples were drawn. It is a mixture of two Gaussians.
- When Δ is very small (top), the resulting density is quite spiky and hallucinates a lot of structure not present in $p(x)$.



Histogram Density as a Function of Bin Width

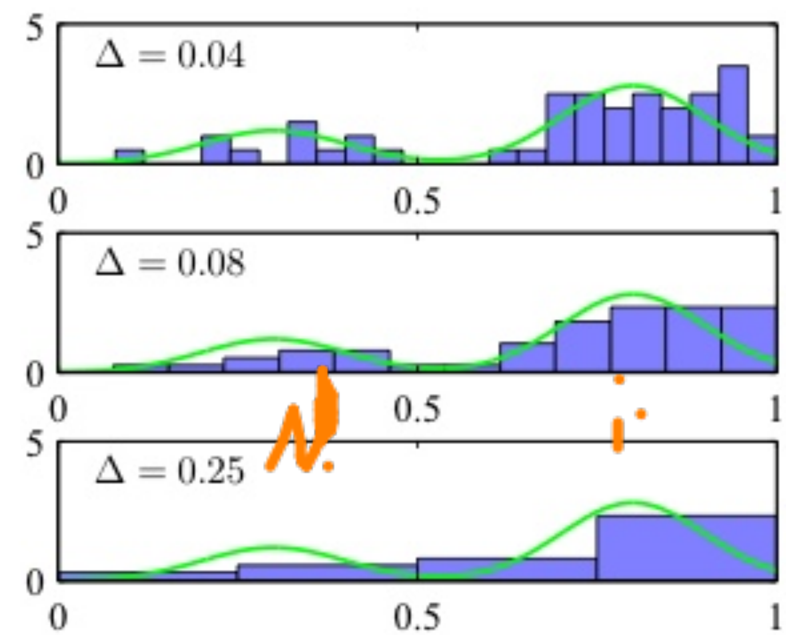
- The green curve is the underlying true density from which the samples were drawn. It is a mixture of two Gaussians.
- When Δ is very small (top), the resulting density is quite spiky and hallucinates a lot of structure not present in $p(x)$.



- When Δ is very big (bottom), the resulting density is quite smooth and consequently fails to capture the bimodality of $p(x)$.

Histogram Density as a Function of Bin Width

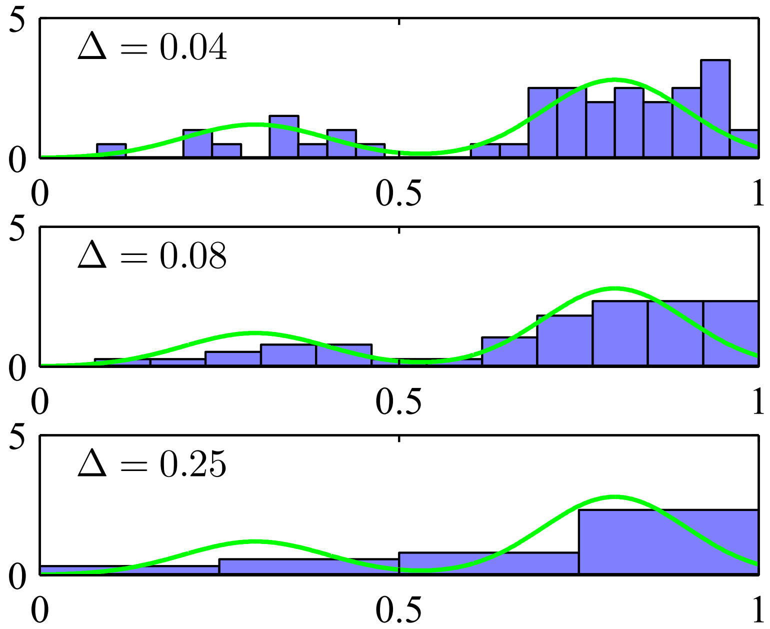
- The green curve is the underlying true density from which the samples were drawn. It is a mixture of two Gaussians.
- When Δ is very small (top), the resulting density is quite spiky and hallucinates a lot of structure not present in $p(x)$.



- When Δ is very big (bottom), the resulting density is quite smooth and consequently fails to capture the bimodality of $p(x)$.
- It appears that the *best results* are obtained for some intermediate value of Δ, which is given in the middle figure.

Histogram Density as a Function of Bin Width

- The green curve is the underlying true density from which the samples were drawn. It is a mixture of two Gaussians.
- When Δ is very small (top), the resulting density is quite spiky and hallucinates a lot of structure not present in $p(x)$.



- When Δ is very big (bottom), the resulting density is quite smooth and consequently fails to capture the bimodality of $p(x)$.
- It appears that the *best results* are obtained for some intermediate value of Δ, which is given in the middle figure.
- In principle, a histogram density model is also dependent on the choice of the edge location of each bin.

Analyzing the Histogram Density

- What are the advantages and disadvantages of the histogram density estimator?

Analyzing the Histogram Density

- What are the advantages and disadvantages of the histogram density estimator?
- Advantages:
  - Simple to evaluate and simple to use.
  - One can throw away $\mathcal{D}$ once the histogram is computed.
  - Can be computed sequentially if data continues to come in.

Analyzing the Histogram Density

- What are the advantages and disadvantages of the histogram density estimator?
- Advantages:
  - Simple to evaluate and simple to use.
  - One can throw away $\mathcal{D}$ once the histogram is computed.
  - Can be computed sequentially if data continues to come in.
- Disadvantages:
  - The estimated density has discontinuities due to the bin edges rather than any property of the underlying density.
  - Scales poorly (curse of dimensionality): we would have M^D bins if we divided each variable in a D-dimensional space into M bins.

What can we learn from Histogram Density Estimation?

- Lesson 1: To estimate the probability density at a particular location, we should consider the data points that lie within some local neighborhood of that point.
  - This requires we define some distance measure.
  - There is a natural smoothness parameter describing the spatial extent of the regions (this was the bin width for the histograms).

What can we learn from Histogram Density Estimation?

- Lesson 1: To estimate the probability density at a particular location, we should consider the data points that lie within some local neighborhood of that point.
  - This requires we define some distance measure.
  - There is a natural smoothness parameter describing the spatial extent of the regions (this was the bin width for the histograms).
- Lesson 2: The value of the smoothing parameter should neither be too large or too small in order to obtain good results.

What can we learn from Histogram Density Estimation?

- Lesson 1: To estimate the probability density at a particular location, we should consider the data points that lie within some local neighborhood of that point.
  - This requires we define some distance measure.
  - There is a natural smoothness parameter describing the spatial extent of the regions (this was the bin width for the histograms).
- Lesson 2: The value of the smoothing parameter should neither be too large or too small in order to obtain good results.
- With these two lessons in mind, we proceed to kernel density estimation and nearest neighbor density estimation, two closely related methods for density estimation.

The Space-Averaged / Smoothed Density

- Consider again samples $\mathbf{x}$ from underlying density $p(\mathbf{x})$.
- Let $\mathcal{R}$ denote a small region containing $\mathbf{x}$.

The Space-Averaged / Smoothed Density

- Consider again samples $\mathbf{x}$ from underlying density $p(\mathbf{x})$.
- Let $\mathcal{R}$ denote a small region containing $\mathbf{x}$.
- The probability mass associated with $\mathcal{R}$ is given by

$$P = \int_{\mathcal{R}} p(\mathbf{x}') d\mathbf{x}' \tag{2}$$

The Space-Averaged / Smoothed Density

- Consider again samples $\mathbf{x}$ from underlying density $p(\mathbf{x})$.
- Let $\mathcal{R}$ denote a small region containing $\mathbf{x}$.
- The probability mass associated with $\mathcal{R}$ is given by

$$P = \int_{\mathcal{R}} p(\mathbf{x}')d\mathbf{x}' \tag{2}$$

- Suppose we have n samples $\mathbf{x} \in \mathcal{D}$. The probability of each sample falling into $\mathcal{R}$ is P.
- How will the total number of k points falling into $\mathcal{R}$ be distributed?

The Space-Averaged / Smoothed Density

- Consider again samples $\mathbf{x}$ from underlying density $p(\mathbf{x})$.
- Let $\mathcal{R}$ denote a small region containing $\mathbf{x}$.
- The probability mass associated with $\mathcal{R}$ is given by

$$P = \int_{\mathcal{R}} p(\mathbf{x}')d\mathbf{x}' \tag{2}$$

- Suppose we have n samples $\mathbf{x} \in \mathcal{D}$. The probability of each sample falling into $\mathcal{R}$ is P.
- How will the total number of k points falling into $\mathcal{R}$ be distributed?
- This will be a **binomial distribution**:

$$P_k = \binom{n}{k} P^k (1-P)^{n-k} \tag{3}$$

The Space-Averaged / Smoothed Density

- The expected value for k is thus

$$\mathcal{E}[k] = nP \tag{4}$$

The Space-Averaged / Smoothed Density

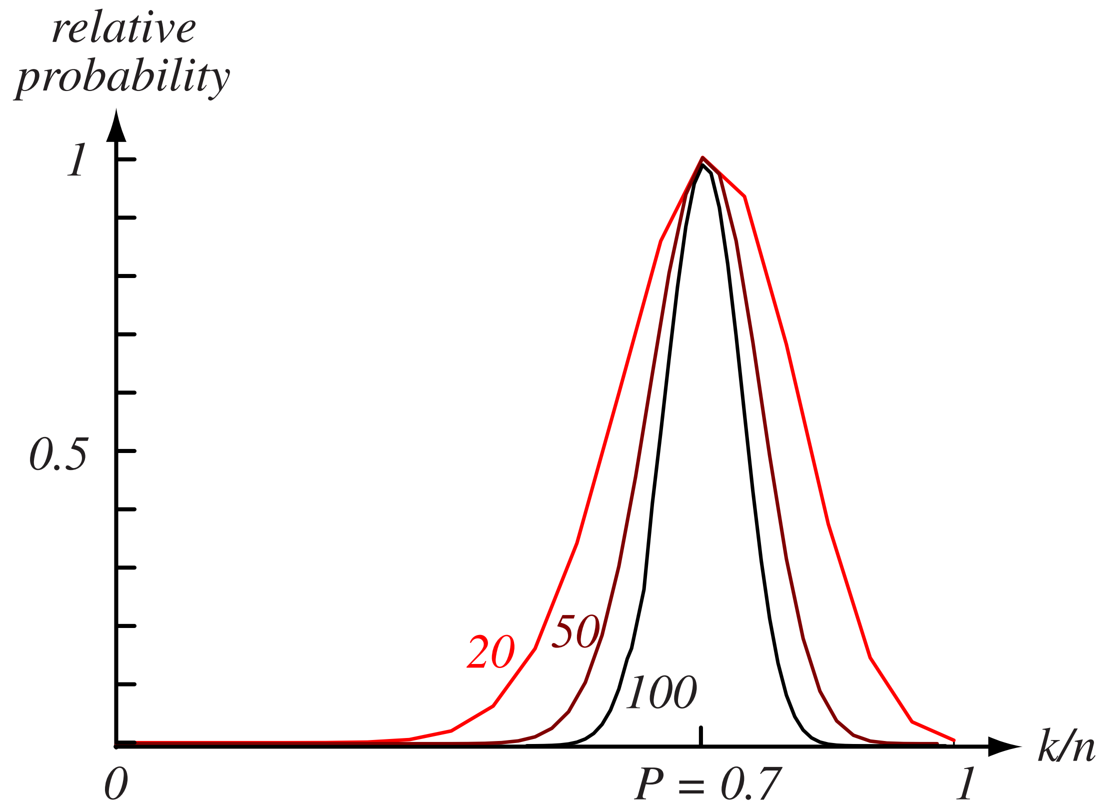
- The expected value for k is thus
$$\mathcal{E}[k] = nP \tag{4}$$
- The binomial for k peaks very sharply about the mean. So, we expect k/n to be a very good estimate for the probability P (and thus for the space-averaged density).

The Space-Averaged / Smoothed Density

- The expected value for k is thus

$$\mathcal{E}[k] = nP \tag{4}$$

- The binomial for k peaks very sharply about the mean. So, we expect k/n to be a very good estimate for the probability P (and thus for the space-averaged density).
- This estimate is increasingly accurate as n increases.



The Space-Averaged / Smoothed Density

- Assuming continuous $p(\mathbf{x})$ and that $\mathcal{R}$ is so small that $p(\mathbf{x})$ does not appreciably vary within it, we can write:

$$\int_{\mathcal{R}} p(\mathbf{x}')d\mathbf{x}' \simeq p(\mathbf{x})V \tag{5}$$

where $\mathbf{x}$ is a point within $\mathcal{R}$ and V is the volume enclosed by $\mathcal{R}$.

The Space-Averaged / Smoothed Density

- Assuming continuous $p(\mathbf{x})$ and that $\mathcal{R}$ is so small that $p(\mathbf{x})$ does not appreciably vary within it, we can write:

$$\int_{\mathcal{R}} p(\mathbf{x}')d\mathbf{x}' \simeq p(\mathbf{x})V \tag{5}$$

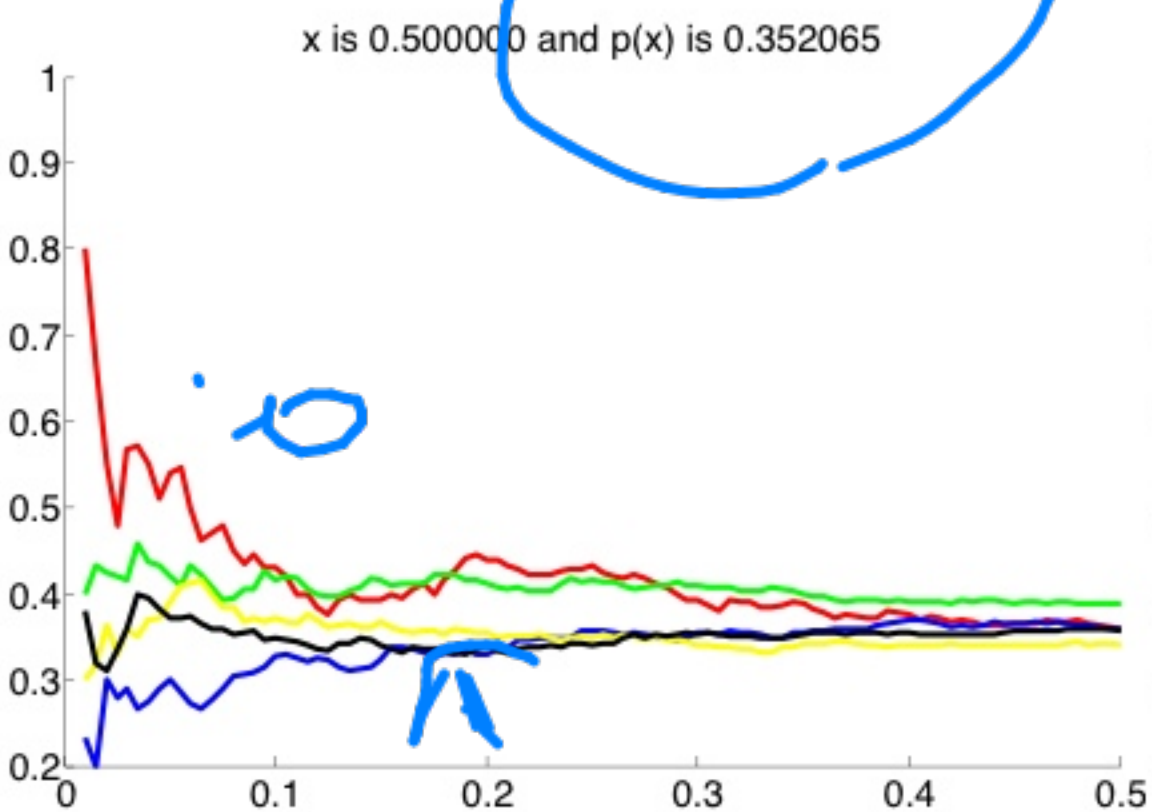
  where $\mathbf{x}$ is a point within $\mathcal{R}$ and V is the volume enclosed by $\mathcal{R}$.

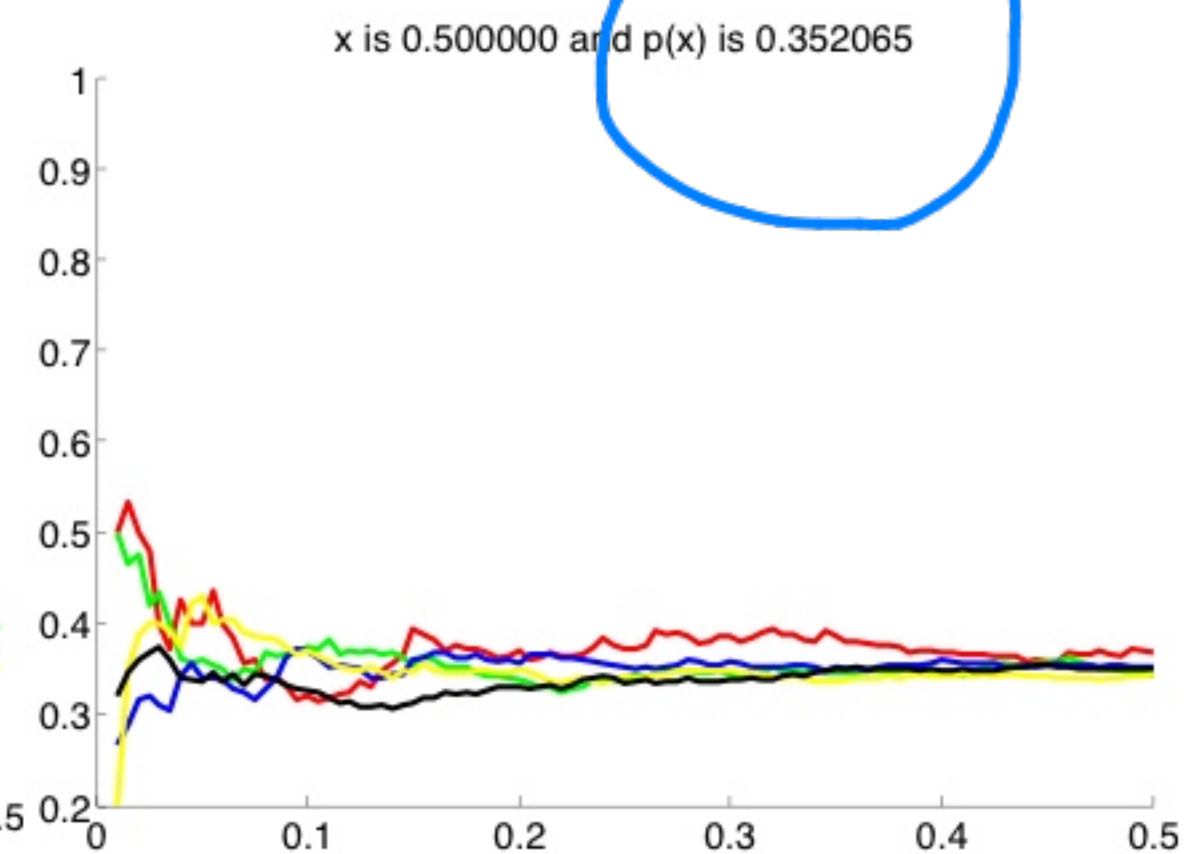
- After some rearranging, we get the following estimate for $p(\mathbf{x})$

$$p(\mathbf{x}) \simeq \frac{k}{nV} \qquad \frac{n_i}{N\Delta_i} \tag{6}$$

Example

- Simulated an example of example the density at 0.5 for an underlying zero-mean, unit variance Gaussian.
- Varied the volume used to estimate the density.
- Red=1000, Green=2000, Blue=3000, Yellow=4000, Black=5000.





Practical Concerns

- The validity of our estimate depends on two contradictory assumptions:
  1. The region $\mathcal{R}$ must be sufficiently small the the density is approximately constant over the region.
  2. The region $\mathcal{R}$ must be sufficiently large that the number k of points falling inside it is sufficient to yield a sharply peaked binomial.

Practical Concerns

- The validity of our estimate depends on two contradictory assumptions:
  1. The region $\mathcal{R}$ must be sufficiently small the the density is approximately constant over the region.
  2. The region $\mathcal{R}$ must be sufficiently large that the number k of points falling inside it is sufficient to yield a sharply peaked binomial.
- Another way of looking it is to fix the volume V and increase the number of training samples. Then, the ratio k/n will converge as desired. But, this will only yield an estimate of the space-averaged density (P/V).

Practical Concerns

- The validity of our estimate depends on two contradictory assumptions:
  1. The region $\mathcal{R}$ must be sufficiently small the the density is approximately constant over the region.
  2. The region $\mathcal{R}$ must be sufficiently large that the number k of points falling inside it is sufficient to yield a sharply peaked binomial.
- Another way of looking it is to fix the volume V and increase the number of training samples. Then, the ratio k/n will converge as desired. But, this will only yield an estimate of the space-averaged density (P/V).
- We want $p(\mathbf{x})$, so we need to let V approach 0. However, with a fixed n, $\mathcal{R}$ will become so small, that no points will fall into it and our estimate would be useless: $p(\mathbf{x}) \simeq 0$.

Practical Concerns

- The validity of our estimate depends on two contradictory assumptions:
  1. The region $\mathcal{R}$ must be sufficiently small the the density is approximately constant over the region.
  2. The region $\mathcal{R}$ must be sufficiently large that the number k of points falling inside it is sufficient to yield a sharply peaked binomial.
- Another way of looking it is to fix the volume V and increase the number of training samples. Then, the ratio k/n will converge as desired. But, this will only yield an estimate of the space-averaged density (P/V).
- We want $p(\mathbf{x})$, so we need to let V approach 0. However, with a fixed n, $\mathcal{R}$ will become so small, that no points will fall into it and our estimate would be useless: $p(\mathbf{x}) \simeq 0$.
- Note that in practice, we cannot let V to become arbitrarily small because the number of samples is always limited.

How can we skirt these limitations when an unlimited number of samples if available?

- To estimate the density at $\mathbf{x}$, form a sequence of regions $\mathcal{R}_1, \mathcal{R}_2, \ldots$ containing $\mathbf{x}$ with the $\mathcal{R}_1$ having 1 sample, $\mathcal{R}_2$ having 2 samples and so on.

How can we skirt these limitations when an unlimited number of samples if available?

- To estimate the density at $\mathbf{x}$, form a sequence of regions $\mathcal{R}_1, \mathcal{R}_2, \ldots$ containing $\mathbf{x}$ with the $\mathcal{R}_1$ having 1 sample, $\mathcal{R}_2$ having 2 samples and so on.
- Let V_n be the volume of $\mathcal{R}_n$, k_n be the number of samples falling in $\mathcal{R}_n$, and $p_n(\mathbf{x})$ be the nth estimate for $p(\mathbf{x})$:

$$p_n(\mathbf{x}) = \frac{k_n}{nV_n} \tag{7}$$

How can we skirt these limitations when an unlimited number of samples if available?

- To estimate the density at $\mathbf{x}$, form a sequence of regions $\mathcal{R}_1, \mathcal{R}_2, \ldots$ containing $\mathbf{x}$ with the $\mathcal{R}_1$ having 1 sample, $\mathcal{R}_2$ having 2 samples and so on.
- Let V_n be the volume of $\mathcal{R}_n$, k_n be the number of samples falling in $\mathcal{R}_n$, and $p_n(\mathbf{x})$ be the nth estimate for $p(\mathbf{x})$:

$R_1 \Rightarrow n = 1$

$R_2 \Rightarrow n = 2$

$$p_n(\mathbf{x}) = \frac{k_n}{nV_n} \tag{7}$$

- If $p_n(\mathbf{x})$ is to converge to $p(\mathbf{x})$ we need the following three conditions

$$\lim_{n\to\infty} V_n = 0 \tag{8}$$

$$\lim_{n\to\infty} k_n = \infty \tag{9}$$

$$\lim_{n\to\infty} k_n/n = 0 \tag{10}$$

- $\lim_{n\to\infty} V_n = 0$ ensures that our space-averaged density will converge to $p(\mathbf{x})$.

- $\lim_{n\to\infty} V_n = 0$ ensures that our space-averaged density will converge to $p(\mathbf{x})$.
- $\lim_{n\to\infty} k_n = \infty$ basically ensures that the frequency ratio will converge to the probability P (the binomial will be sufficiently peaked).

- $\lim_{n\to\infty} V_n = 0$ ensures that our space-averaged density will converge to $p(\mathbf{x})$.
- $\lim_{n\to\infty} k_n = \infty$ basically ensures that the frequency ratio will converge to the probability P (the binomial will be sufficiently peaked).
- $\lim_{n\to\infty} k_n/n = 0$ is required for $p_n(\mathbf{x})$ to converge at all. It also says that although a huge number of samples will fall within the region $\mathcal{R}_n$, they will form a negligibly small fraction of the total number of samples.

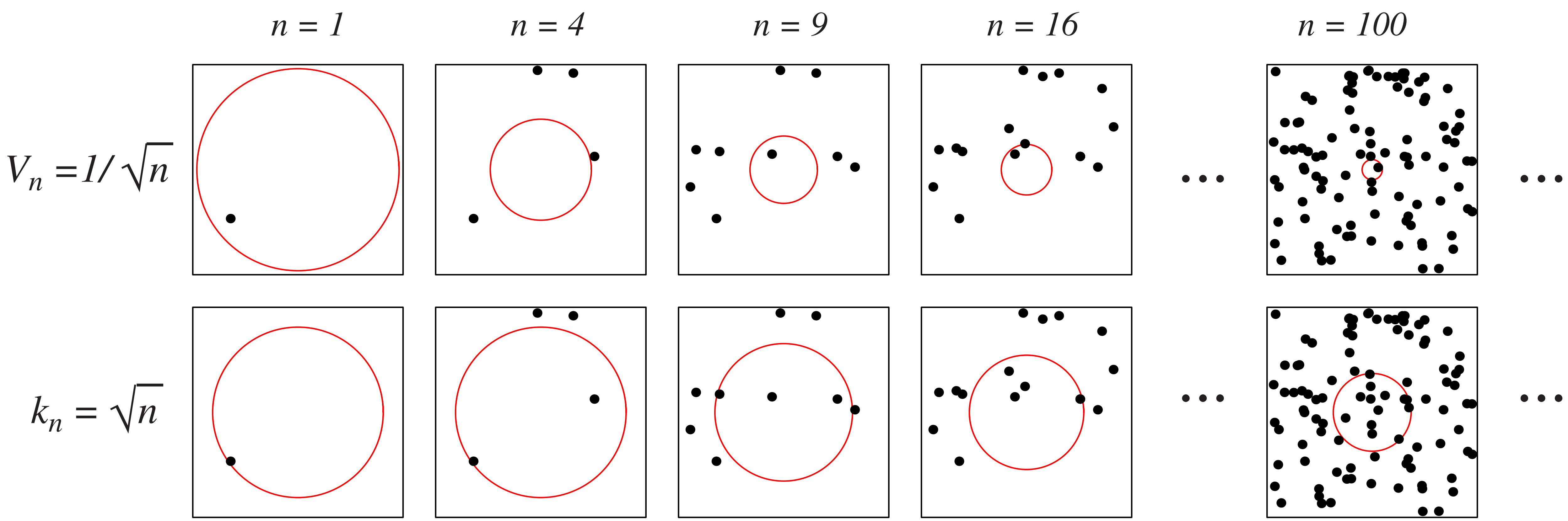
- $\lim_{n\to\infty} V_n = 0$ ensures that our space-averaged density will converge to $p(\mathbf{x})$.
- $\lim_{n\to\infty} k_n = \infty$ basically ensures that the frequency ratio will converge to the probability P (the binomial will be sufficiently peaked).
- $\lim_{n\to\infty} k_n/n = 0$ is required for $p_n(\mathbf{x})$ to converge at all. It also says that although a huge number of samples will fall within the region $\mathcal{R}_n$, they will form a negligibly small fraction of the total number of samples.
- There are two common ways of obtaining regions that satisfy these conditions:

- $\lim_{n\to\infty} V_n = 0$ ensures that our space-averaged density will converge to $p(\mathbf{x})$.
- $\lim_{n\to\infty} k_n = \infty$ basically ensures that the frequency ratio will converge to the probability P (the binomial will be sufficiently peaked).
- $\lim_{n\to\infty} k_n/n = 0$ is required for $p_n(\mathbf{x})$ to converge at all. It also says that although a huge number of samples will fall within the region $\mathcal{R}_n$, they will form a negligibly small fraction of the total number of samples.
- There are two common ways of obtaining regions that satisfy these conditions:
  1. Shrink an initial region by specifying the volume V_n as some function of n such as $V_n = 1/\sqrt{n}$. Then, we need to show that $p_n(\mathbf{x})$ converges to $p(\mathbf{x})$. (This is like the Parzen window we'll talk about next.)

- $\lim_{n\to\infty} V_n = 0$ ensures that our space-averaged density will converge to $p(\mathbf{x})$.
- $\lim_{n\to\infty} k_n = \infty$ basically ensures that the frequency ratio will converge to the probability P (the binomial will be sufficiently peaked).
- $\lim_{n\to\infty} k_n/n = 0$ is required for $p_n(\mathbf{x})$ to converge at all. It also says that although a huge number of samples will fall within the region $\mathcal{R}_n$, they will form a negligibly small fraction of the total number of samples.
- There are two common ways of obtaining regions that satisfy these conditions:
  1. Shrink an initial region by specifying the volume V_n as some function of n such as $V_n = 1/\sqrt{n}$. Then, we need to show that $p_n(\mathbf{x})$ converges to $p(\mathbf{x})$. (This is like the Parzen window we'll talk about next.)
  2. Specify k_n as some function of n such as $k_n = \sqrt{n}$. Then, we grow the volume V_n until it encloses k_n neighbors of $\mathbf{x}$. (This is the k-nearest-neighbor).

- $\lim_{n\to\infty} V_n = 0$ ensures that our space-averaged density will converge to $p(\mathbf{x})$.
- $\lim_{n\to\infty} k_n = \infty$ basically ensures that the frequency ratio will converge to the probability P (the binomial will be sufficiently peaked).
- $\lim_{n\to\infty} k_n/n = 0$ is required for $p_n(\mathbf{x})$ to converge at all. It also says that although a huge number of samples will fall within the region $\mathcal{R}_n$, they will form a negligibly small fraction of the total number of samples.
- There are two common ways of obtaining regions that satisfy these conditions:
  1. Shrink an initial region by specifying the volume V_n as some function of n such as $V_n = 1/\sqrt{n}$. Then, we need to show that $p_n(\mathbf{x})$ converges to $p(\mathbf{x})$. (This is like the Parzen window we'll talk about next.)
  2. Specify k_n as some function of n such as $k_n = \sqrt{n}$. Then, we grow the volume V_n until it encloses k_n neighbors of $\mathbf{x}$. (This is the k-nearest-neighbor).

  Both of these methods converge...



$n = 1$
$n = 4$
$n = 9$
$n = 16$
$n = 100$
$V_n = 1/\sqrt{n}$
$k_n = \sqrt{n}$