Estimating probability densities

- Thus far we’ve only discussed probability
Estimating probability densities

- Thus far we’ve only discussed probability
- We haven’t yet touched on the inductive leap from data to probabilities underlying the data
Estimating probability densities

- Thus far we’ve only discussed probability
- We haven’t yet touched on the inductive leap from data to probabilities underlying the data
- That is a large part of what STATISTICAL INFERENCE is
Estimating probability densities

- Thus far we’ve only discussed probability
- We haven’t yet touched on the inductive leap from data to probabilities underlying the data
- That is a large part of what STATISTICAL INFERENCE is
- We’re now ready to take our first foray into this: estimating PROBABILITY DENSITIES
Estimating probability densities

▸ Thus far we’ve only discussed probability
▸ We haven’t yet touched on the inductive leap from data to probabilities underlying the data
▸ That is a large part of what STATISTICAL INFERENCE is
▸ We’re now ready to take our first foray into this: ESTIMATING PROBABILITY DENSITIES
▸ We’ll distinguish an estimated probability distribution from an underlying distribution with the \( \hat{P} \) symbol, so that we have \( \hat{P} \) estimating \( P \)
Estimating probability densities

- Thus far we’ve only discussed probability
- We haven’t yet touched on the inductive leap from data to probabilities underlying the data
- That is a large part of what STATISTICAL INference is
- We’re now ready to take our first foray into this: ESTIMATING PROBABILITY DENSITIES
- We’ll distinguish an estimated probability distribution from an underlying distribution with the $\hat{\cdot}$ symbol, so that we have $\hat{P}$ estimating $P$
- We’ll distinguish estimated from true parameter values the same way: e.g., $\hat{\pi}$ vs. $\pi$
Estimating discrete densities

- English BINOMIALS:

  principal and interest     interest and principal
  ice and snow               snow and ice
  black and white            white and black
Estimating discrete densities

- **English BINOMIALS:**
  
  \[
  \begin{array}{ll}
  \text{principal and interest} & \text{interest and principal} \\
  \text{ice and snow} & \text{snow and ice} \\
  \text{black and white} & \text{white and black}
  \end{array}
  \]

- Suppose we’re interested in estimating the Bernoulli parameter $\pi$ associated with the ordering preference of a specific English binomial, $\{\text{interest, principal}\}$
Estimating discrete densities

- English BINOMIALS:
  
  \[ \text{principal and interest} \quad \text{interest and principal} \]
  \[ \text{ice and snow} \quad \text{snow and ice} \]
  \[ \text{black and white} \quad \text{white and black} \]

- Suppose we’re interested in estimating the Bernoulli parameter \( \pi \) associated with the ordering preference of a specific English binomial, \( \{ \text{interest, principal} \} \)
  - Arbitrarily call \text{principal and interest} “success”
Estimating discrete densities

- **English BINOMIALS:**
  
  - *principal and interest*  *interest and principal*
  - *ice and snow*  *snow and ice*
  - *black and white*  *white and black*

- Suppose we’re interested in estimating the Bernoulli parameter $\pi$ associated with the ordering preference of a specific English binomial, \{*interest*,*principal*\}
  
  - Arbitrarily call *principal and interest* “success”

- Say we observe the binomial 20 times:
  
  - *principal and interest* 14
  - *interest and principal* 6
Estimating discrete densities

- **English BINOMIALS:**
  
  \[
  \begin{align*}
  principal \ and \ interest & \quad interest \ and \ principal \\
  ice \ and \ snow & \quad snow \ and \ ice \\
  black \ and \ white & \quad white \ and \ black
  \end{align*}
  \]

- Suppose we’re interested in estimating the Bernoulli parameter \( \pi \) associated with the ordering preference of a specific English binomial, \( \{interest,principal\} \)
  - Arbitrarily call **principal and interest** “success”

- Say we observe the binomial 20 times:
  
  \[
  \begin{align*}
  principal \ and \ interest & \quad 14 \\
  interest \ and \ principal & \quad 6
  \end{align*}
  \]

- A clearly intuitive way of choosing \( \hat{\pi} \):
  
  \[
  binomial parameter = \frac{14}{14 + 6}
  \]
Estimating discrete densities

- English binomials:
  
  principal and interest  interest and principal
  ice and snow           snow and ice
  black and white        white and black

- Suppose we’re interested in estimating the Bernoulli parameter $\pi$ associated with the ordering preference of a specific English binomial, $\{\text{interest,principal}\}$
  - Arbitrarily call principal and interest “success”

- Say we observe the binomial 20 times:
  
  principal and interest  14
  interest and principal  6

- A clearly intuitive way of choosing $\hat{\pi}$:
  
  \[
  \text{binomial parameter} = \frac{14}{14 + 6}
  \]
Estimating discrete densities

- English BINOMIALS:
  
  *principal and interest*  
  *interest and principal*  
  *ice and snow*  
  *snow and ice*  
  *black and white*  
  *white and black*  

- Suppose we’re interested in estimating the Bernoulli parameter $\pi$ associated with the ordering preference of a specific English binomial, $\{\text{interest, principal}\}$
  
  - Arbitrarily call *principal and interest* “success”

- Say we observe the binomial 20 times:
  
  *principal and interest* 14  
  *interest and principal* 6  

- A clearly intuitive way of choosing $\hat{\pi}$:
  
  $$binomial parameter = \frac{14}{14 + 6}$$
Estimating discrete densities

- **English BINOMIALS:**

  \[ \begin{align*}
  \text{principal and interest} & \quad \text{interest and principal} \\
  \text{ice and snow} & \quad \text{snow and ice} \\
  \text{black and white} & \quad \text{white and black}
  \end{align*} \]

- Suppose we’re interested in estimating the Bernoulli parameter \( \pi \) associated with the ordering preference of a specific English binomial, \( \{ \text{interest, principal} \} \)
  - Arbitrarily call \( \text{principal and interest} \) “success”

- Say we observe the binomial 20 times:

  \[ \begin{align*}
  \text{principal and interest} & = 14 \\
  \text{interest and principal} & = 6
  \end{align*} \]

- A clearly intuitive way of choosing \( \hat{\pi} \):

  \[
  \text{binomial parameter} = \frac{14}{14 + 6} = \frac{14}{20} = 0.7
  \]

- This is called RELATIVE FREQUENCY ESTIMATION!
For multinomial distributions

- This method extends to multinomial distributions

<table>
<thead>
<tr>
<th></th>
<th>⟨SB, DO⟩</th>
<th>⟨SB, IO⟩</th>
<th>⟨DO, SB⟩</th>
<th>⟨DO, IO⟩</th>
<th>⟨IO, SB⟩</th>
<th>⟨IO, DO⟩</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count</td>
<td>478</td>
<td>59</td>
<td>1</td>
<td>3</td>
<td>20</td>
<td>9</td>
</tr>
<tr>
<td>Relative Freq.</td>
<td>0.839</td>
<td>0.104</td>
<td>0.001</td>
<td>0.005</td>
<td>0.035</td>
<td>0.016</td>
</tr>
</tbody>
</table>
For multinomial distributions

- This method extends to multinomial distributions
- e.g., German NP pairs:

  Mir gefällt der Film nicht.
  me.DAT like the film.NOM not.

<table>
<thead>
<tr>
<th></th>
<th>⟨SB, DO⟩</th>
<th>⟨SB, IO⟩</th>
<th>⟨DO, SB⟩</th>
<th>⟨DO, IO⟩</th>
<th>⟨IO, SB⟩</th>
<th>⟨IO, DO⟩</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count</td>
<td>478</td>
<td>59</td>
<td>1</td>
<td>3</td>
<td>20</td>
<td>9</td>
</tr>
<tr>
<td>Relative Freq.</td>
<td>0.839</td>
<td>0.104</td>
<td>0.001</td>
<td>0.005</td>
<td>0.035</td>
<td>0.016</td>
</tr>
</tbody>
</table>
Relative frequency estimation: to estimate the parameters of a $k$-class multinomial distribution from count data where the observed counts in each class are $n_1, n_2, \ldots, n_k$, set

$$\hat{\pi}_i = \frac{n_i}{N}$$

where $N = \sum_{i=1}^{k}$. 
Relative frequency estimation

**Relative frequency estimation:** to estimate the parameters of a $k$-class multinomial distribution from count data where the observed counts in each class are $n_1, n_2, \ldots, n_k$, set

$$\hat{\pi}_i = \frac{n_i}{N}$$

where $N = \sum_{i=1}^{k}$. 

- This is a highly general-purpose method for estimating the parameters of multinomial distributions
Relative frequency estimation: to estimate the parameters of a $k$-class multinomial distribution from count data where the observed counts in each class are $n_1, n_2, \ldots, n_k$, set

$$\hat{\pi}_i = \frac{n_i}{N}$$

where $N = \sum_{i=1}^{k} n_i$.

- This is a highly general-purpose method for estimating the parameters of multinomial distributions
- It has advantages and disadvantages (think about these; we’ll return to them in Chapter 4)
Here are F0 formant frequencies for adult male English speakers speaking the vowel [α]
Estimating continuous densities

- Here are F0 formant frequencies for adult male English speakers speaking the vowel [a].

Clearly we can’t do relative frequency estimation with these data!
Histograms for continuous density estimation

- One option: the HISTOGRAM
Histograms for continuous density estimation

- One option: the **HISTOGRAM**
- This is something you’re probably familiar with:
Histograms for continuous density estimation

- One option: the **HISTOGRAM**
- This is something you’re probably familiar with:

![Histogram Graph]

- What is it that we’re doing when we construct a histogram?
Histogrm for continuous density estimation

To construct a histogram:
Histograms for continuous density estimation

To construct a histogram:

- Divide some part of the range of your random variable $X$ into $K$ equally-spaced bins of width $w$;
Histograms for continuous density estimation

To construct a histogram:

- Divide some part of the range of your random variable $X$ into $K$ equally-spaced bins of width $w$;
- Count the number of observations $n_i$ that fall into each bin $1, \ldots, i, \ldots, k$
Histograms for continuous density estimation

To construct a histogram:

- Divide some part of the range of your random variable $X$ into $K$ equally-spaced bins of width $w$;
- Count the number of observations $n_i$ that fall into each bin $1, \ldots, i, \ldots, k$;
- Use relative frequency estimation to estimate the probabilities of each bin $P(c_i)$.
Histograms for continuous density estimation

To construct a histogram:

- Divide some part of the range of your random variable \( X \) into \( K \) equally-spaced bins of width \( w \);
- Count the number of observations \( n_i \) that fall into each bin \( 1, \ldots, i, \ldots, k \);
- Use relative frequency estimation to estimate the probabilities of each bin \( P(c_i) \);
- Assume a uniform distribution within the bin.
Histograms for continuous density estimation

To construct a histogram:

- Divide some part of the range of your random variable $X$ into $K$ equally-spaced bins of width $w$;
- Count the number of observations $n_i$ that fall into each bin $1, \ldots, i, \ldots, k$;
- Use relative frequency estimation to estimate the probabilities of each bin $P(c_i)$;
- Assume a uniform distribution within the bin;
- Thus we get a distribution of the form

$$ P(X = x) = \begin{cases} \frac{n_i}{Nw} & \text{if } x \text{ is in the } i\text{-th bin} \\ 0 & \text{otherwise} \end{cases} $$
Disadvantages of histograms

- it can assign zero probability to intervals for which the data seem to suggest possible outcomes
Disadvantages of histograms

- it can assign zero probability to intervals for which the data seem to suggest possible outcomes
- the shape of the histogram can be sensitive to the exact positioning of the bin
An alternative: kernel density estimation

- The intuition: new observations are likely to be similar to old observations, but will generally be a bit different
An alternative: kernel density estimation

- The intuition: new observations are likely to be similar to old observations, but will generally be a bit different
- So for each observation, put some probability mass in the “neighborhood” of the observation
An alternative: kernel density estimation

- The intuition: new observations are likely to be similar to old observations, but will generally be a bit different
- So for each observation, put some probability mass in the “neighborhood” of the observation
- Sum up the probability masses you put down this way to get a continuous density estimate
An alternative: kernel density estimation

- The intuition: new observations are likely to be similar to old observations, but will generally be a bit different.
- So for each observation, put some probability mass in the “neighborhood” of the observation.
- Sum up the probability masses you put down this way to get a continuous density estimate.
An alternative: kernel density estimation

- The intuition: new observations are likely to be similar to old observations, but will generally be a bit different
- So for each observation, put some probability mass in the “neighborhood” of the observation
- Sum up the probability masses you put down this way to get a continuous density estimate
An alternative: kernel density estimation

- The intuition: new observations are likely to be similar to old observations, but will generally be a bit different
- So for each observation, put some probability mass in the “neighborhood” of the observation
- Sum up the probability masses you put down this way to get a continuous density estimate
Kernel Density Estimation

- a kernel $K$ takes an observation $x_i$ and returns a non-negative function $K(x_i, \cdot)$ which distributes a total probability mass of 1 over the range of possible outcomes. Hence

$$
\int_x K(x_i, x) \, dx = 1
$$

(properness).
Kernel Density Estimation

- A kernel $K$ takes an observation $x_i$ and returns a non-negative function $K(x_i, \cdot)$ which distributes a total probability mass of 1 over the range of possible outcomes. Hence

$$\int_K K(x_i, x) \, dx = 1$$

(properness).

- The kernel density estimate is the average of the assigned densities:

$$\hat{p}(x) = \frac{1}{n} \sum_{i=1}^{n} K(x_i, x)$$
Kernel Density Estimation

- a kernel \( K \) takes an observation \( x_i \) and returns a non-negative function \( K(x_i, \cdot) \) which distributes a total probability mass of 1 over the range of possible outcomes. Hence

\[
\int_x K(x_i, x) \, dx = 1
\]

(properness).

- The kernel density estimate is the average of the assigned densities:

\[
\hat{p}(x) = \frac{1}{n} \sum_{i=1}^{n} K(x_i, x)
\]

- On the previous slide, we saw the normal or Gaussian kernel:

\[
\hat{p}(X = x) = \frac{1}{n \sqrt{2\pi} b^2} \sum_{i=1}^{n} \exp \left[ -\frac{(x - x_i)^2}{2b^2} \right]
\]
Kernel Density Estimation

- A kernel $K$ takes an observation $x_i$ and returns a non-negative function $K(x_i, \cdot)$ which distributes a total probability mass of 1 over the range of possible outcomes. Hence

$$\int_{x} K(x_i, x) \, dx = 1$$

(properness).

- The kernel density estimate is the average of the assigned densities:

$$\hat{p}(x) = \frac{1}{n} \sum_{i=1}^{n} K(x_i, x)$$

- On the previous slide, we saw the normal or Gaussian kernel:

$$\hat{p}(X = x) = \frac{1}{n\sqrt{2\pi}b^2} \sum_{i=1}^{n} \exp \left[ -\frac{(x - x_i)^2}{2b^2} \right]$$
Kernel Density Estimation

- A kernel $K$ takes an observation $x_i$ and returns a non-negative function $K(x_i, \cdot)$ which distributes a total probability mass of 1 over the range of possible outcomes. Hence

$$\int_{x} K(x_i, x) \, dx = 1$$

(properness).

- The kernel density estimate is the average of the assigned densities:

$$\hat{p}(x) = \frac{1}{n} \sum_{i=1}^{n} K(x_i, x)$$

- On the previous slide, we saw the normal or Gaussian kernel:

$$\hat{p}(X = x) = \frac{1}{n \sqrt{2\pi b^2}} \sum_{i=1}^{n} \exp \left[ -\frac{(x - x_i)^2}{2b^2} \right]$$

Compare this with the normal distribution:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[ -\frac{(x - \mu)^2}{2\sigma^2} \right]$$
Kernel Density Estimation

Results of kernel density estimation for our problem, $b = 5$:
Kernel Density Estimation

- The choice of kernel, however, is up to you!
Kernel Density Estimation

- The choice of kernel, however, is up to you!
- Another option: RECTANGULAR KERNEL – put a uniform distribution of width $b$ around each observation!

![Graph showing F0 frequency (Hz) vs. p(x) distribution.]
Kernel Density Estimation

▶ You can even use kernel density estimation for discrete densities!
Kernel Density Estimation

- You can even use kernel density estimation for discrete densities!
- Example: consider a mini-lexicon of CVC English words constructable from the phonemes /t/, /p/, /k/, /æ/, /ʌ/, and /ʊ/
Kernel Density Estimation

- You can even use kernel density estimation for discrete densities!
- Example: consider a mini-lexicon of CVC English words constructable from the phonemes /t/, /p/, /k/, /æ/, /ʌ/, and /ʊ/
- Here are the words of English in that set of possible words:
**Kernel Density Estimation**

- You can even use kernel density estimation for discrete densities!
- Example: consider a mini-lexicon of CVC English words constructable from the phonemes /t/,/p/,/k/,/æ/,/ʌ/, and/ʊ/
- Here are the words of English in that set of possible words:

```
  tʌt  tʌk  tʌp  kʌt  kʌk  kʌp  pʌt  pʌk  pʌp  tæt  tæk  tæp  kæt  kæk  kæp  pæt  pæk  pæp  tʊt  tʊk  tʊp  kʊt  kʊk  kʊp  pʊt  pʊk  pʊp
```

- We might want a gradient model of “English non-words that could have been words” (a PHONOTACTIC model)
Kernel Density Estimation

- You can even use kernel density estimation for discrete densities!
- Example: consider a mini-lexicon of CVC English words constructable from the phonemes /t/, /p/, /k/, /æ/, /ʌ/, and /ʊ/
- Here are the words of English in that set of possible words:

![Graph showing the distribution of possible words]

- We might want a gradient model of “English non-words that could have been words” (a PHONOTACTIC model)
- How could we construct a kernel density estimate over possible words in this space?
Kernel Density Estimation

- You can even use kernel density estimation for discrete densities!
- Example: consider a mini-lexicon of CVC English words constructable from the phonemes /t/,/p/,/k/,/æ/,/ʌ/, and/ʊ/
- Here are the words of English in that set of possible words:

![](image)

- We might want a gradient model of “English non-words that could have been words” (a PHONOTACTIC model)
- How could we construct a kernel density estimate over possible words in this space?
References I