IV. Data Summaries

A. Pictures of Data

D&P: 1.4B

1. Represent Categorical data using a bar plot:
   a. Most easily represented in a table; ex., a measure of number of lifetime episodes of “instrumental” drug use by doctors (N Engl J Med 1986; 315:805-10) (Table 1)

   Table 1: Frequencies of Incidence Counts for Instrumental Drug Use Among Physicians

<table>
<thead>
<tr>
<th>Never</th>
<th>Few</th>
<th>Some</th>
<th>Many</th>
</tr>
</thead>
<tbody>
<tr>
<td>238</td>
<td>41</td>
<td>6</td>
<td>6</td>
</tr>
</tbody>
</table>

   i. The sum of these frequencies is 291
   b. Counts of how often each value is observed are called frequencies.
      i. Can represent in a bar plot; see Fig. 1/ for drug use example.
      ii. Counts of how often each value is observed, divided by total, are called relative frequencies. (Table 2)
   c. Most usefully represented in bar plot.
For example, drug use in Fig. 2/i.

2. Represent Numerical data using a histogram.

   a. Objective: Draw a picture that in some sense captures the broad features of the data, perhaps at the price of leaving finer details obscure.
Fig. 2: Relative Frequencies of Instrumental Drug Use

i. If data represent observations from some population, we hope the same features hold for population.

b. Procedure:
   i. divide range of possible data values into a set of (usually equally-lengthed) \textit{class intervals}.
   ii. Count how many data values fall in each class interval.
   iii. Initially represent number of counts in each class interval by rectangle whose height is the number of items in that class interval.
### Table 3: Log Blood Lead Levels for Renters

<table>
<thead>
<tr>
<th>Interval</th>
<th>Frequency</th>
<th>Relative Frequency</th>
<th>Histogram Height on Density Scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1 \leq \log\text{ Pb} &lt; 1.5)</td>
<td>5</td>
<td>0.132</td>
<td>0.264</td>
</tr>
<tr>
<td>(1.5 \leq \log\text{ Pb} &lt; 2)</td>
<td>14</td>
<td>0.368</td>
<td>0.736</td>
</tr>
<tr>
<td>(2 \leq \log\text{ Pb} &lt; 2.5)</td>
<td>10</td>
<td>0.263</td>
<td>0.526</td>
</tr>
<tr>
<td>(2.5 \leq \log\text{ Pb} &lt; 3)</td>
<td>8</td>
<td>0.211</td>
<td>0.422</td>
</tr>
<tr>
<td>(3 \leq \log\text{ Pb} &lt; 3.5)</td>
<td>0</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>(3.5 \leq \log\text{ Pb} &lt; 4)</td>
<td>1</td>
<td>0.026</td>
<td>0.052</td>
</tr>
</tbody>
</table>

**D&P: 3.3, 3.5**

- Rectangles sit next to each other without separation
- Occasionally a data value will fall exactly on the edge of one of the class interval.
  
  ▶ A convention needed for deciding which class interval to place it into.
  
  ▶ Two choices:
    
    Bin contains items that equal lower limit but not upper limit
    
    Bin contains items that equal upper limit but not lower limit

**c. Example:** Histogram calculations for Renters (Table 3)

  **i.** variables on the boundary belong to lower bin
d. How many intervals?
   i. Too few obscure patterns in the data
   ii. Too many perhaps reveal too much information
       • perhaps too much to understand, at least in a first look
       • so much that what might be chance occurrences are highlighted.

   D&P: 3.2

3. What might one see?
   a. outliers: observations lying far from the rest of the data. For example, ages for people expected to be 65+ could have very low ages mixed in. (Fig. 3/).
      i. Could represent
         • data miscoded; in example, age calculation did not allow for ages over 100; this is corrected (Fig. 4/).
         • a part of a process that we might not be measuring correctly.
         • A rare strong effect that we are interested in.
   b. Aspects of remaining data
      i. Center
      ii. Spread
iii. Shape

- Symmetric?
- Skewed left or right? (direction of skewness is direction of long tail)

iv. Number of peaks, or modes

- unimodal, bimodal, etc.
• Careful: This is the least–certain thing that we might try to read off of the histogram.

4. Problem: Hard to compare 2 histograms with different sample sizes. (Fig. 5/ vs. Fig. 6/).
   a. choice of arbitrary features like sample size shouldn’t change gross features of picture.
b. Solution: Let height of bar be the proportion in the class interval (Fig. 7/ and Fig. 8/).

5. Problem: Height of histogram strongly depends on irrelevant
choice of bin size

a. Rule also says combining 2 class intervals together makes height of combined rectangle the sum of individuals

b. whereas if they are initially the same height they should remain the same height:
c. choice of arbitrary features like class interval width shouldn’t change gross features of picture.

d. Solution: Let height of bar be the proportion in the class interval divided by the length of the interval (Fig. 9/ and Fig. 10/).
Fig. 8: Blood Lead Levels for Renters, Relative Frequency Scale

i. Makes areas represent proportions.

ii. Makes height of class interval formed by collapsing
2 adjacent class interval the avg. of the original heights.

e. Obvious features:
   i. histogram is never negative
   ii. area underneath whole thing is 1

f. Population Analogue:
i. Hope that histogram tells you something about population you are sampling from

ii. For a finite population, it should approximate histogram we
would get if we sampled everyone.

iii. For infinite populations, areas should approximate long run proportions of times we will see observations in interval.

iv. Use density scale.

v. Unlike bar chart for categorical variables, don’t leave space between the bars.