Last time we learned that the shape of a histogram depends on the bin size. Too many breaks, or not enough breaks, both lead to useless histograms. So, “turn the knob” on breaks & explore the shapes.

The shape of a histogram is very important, and conveys lots of info. The interpretation of histograms is an “art” that you will learn through practice. Here is an example:

Grades from spring 17:

Ignore the “small” features. See the “big picture.”

What’s small?

What’s large?

Ans. It depends!

There is a great deal of useful info in a histogram:

- center (location) of data = typical value
- spread of data, = typical spread of values
- shape of data, ... All tell a good story.

In the future, the first thing you should do when you see a bunch of observations (either numbers or not) is histogram them. You will learn something!
Histograms can show frequency, or relative freq. on the y-axis.

Rel. freq. = freq. / total sample size.  (examples on next page).

**For Continuous r.v.** Rel. freq. hists have a very important property that will become useful later (when we get into inference): 

\[
\text{prob}(a < x < b) = \text{proportion of times} \ (a < x < b) = \text{area between} \ a, \ b.
\]

So, shift your focus from the y-axis to area.

This property has a surprising consequence: \( \text{prob}(x = a) = 0 \)!

**Example:** Suppose \( x = \text{cont.} \) and \( \longrightarrow \) Then, \( \text{pr}(x \leq 2.5) = \text{blue area} = \text{pr}(x < 2.5) \)

Intuition: if \( x \) is truly cont. Then the prob of getting exactly 2.500000... is zero.

**For discrete/categ. r.v.**, probs are just sums of proportions for each \( x \):

**E.g.** Suppose \( x = \text{discrete with the following hist.} \) What’s the prob. \( (\text{or prop.}) \) of \( 2 \leq x \leq 3 \)? 80%.

I.e. for discrete/categ. area is not important.
Example of rel-freq. hist and their interpretation:

### Form G: Lecture -- Assignments

<table>
<thead>
<tr>
<th>Question</th>
<th>Excellent</th>
<th>Very Good</th>
<th>Good</th>
<th>Fair</th>
<th>Poor</th>
<th>Very Poor</th>
<th>Median</th>
</tr>
</thead>
<tbody>
<tr>
<td>The course as a whole:</td>
<td>27%</td>
<td>33%</td>
<td>22%</td>
<td>12%</td>
<td>3%</td>
<td>3%</td>
<td>3.80</td>
</tr>
<tr>
<td>Textbook overall:</td>
<td>33%</td>
<td>30%</td>
<td>27%</td>
<td>10%</td>
<td>0%</td>
<td>0%</td>
<td>3.94</td>
</tr>
<tr>
<td>Instructor overall:</td>
<td>50%</td>
<td>28%</td>
<td>10%</td>
<td>7%</td>
<td>2%</td>
<td>3%</td>
<td>4.50</td>
</tr>
<tr>
<td>Instructor's contribution:</td>
<td>42%</td>
<td>27%</td>
<td>15%</td>
<td>8%</td>
<td>3%</td>
<td>3%</td>
<td>4.22</td>
</tr>
<tr>
<td>Instructor's interest:</td>
<td>53%</td>
<td>26%</td>
<td>7%</td>
<td>5%</td>
<td>2%</td>
<td>7%</td>
<td>4.56</td>
</tr>
<tr>
<td>Amount learned:</td>
<td>39%</td>
<td>27%</td>
<td>20%</td>
<td>8%</td>
<td>3%</td>
<td>2%</td>
<td>4.09</td>
</tr>
<tr>
<td>Relevance and usefulness of homework:</td>
<td>37%</td>
<td>17%</td>
<td>27%</td>
<td>12%</td>
<td>3%</td>
<td>3%</td>
<td>3.75</td>
</tr>
</tbody>
</table>

For median calculation: 5 = Excellent 4 = Very Good 3 = Good 2 = Fair 1 = Poor 0 = Very Poor

There are many more of these at the bottom of course website.

The numbers on each row say something about rating, i.e., how the students rated something. So, the random variable is rating.

![Histogram](image)

FREQ x 61 Rel. Freq.

<table>
<thead>
<tr>
<th>Rating</th>
<th>Freq.</th>
<th>Rel. Freq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>20</td>
<td>0.33</td>
</tr>
<tr>
<td>4</td>
<td>16</td>
<td>0.24</td>
</tr>
<tr>
<td>3</td>
<td>13</td>
<td>0.22</td>
</tr>
<tr>
<td>2</td>
<td>7</td>
<td>0.11</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>0.03</td>
</tr>
<tr>
<td>0</td>
<td>2</td>
<td>0.03</td>
</tr>
</tbody>
</table>

The median is 3.80, so median (and mean) make sense only for quantitative data.

Interpret: center ~ 3 or 4 spread ~ 1, 1.5 shape ~ skewed (to ...)

Also, can find prob(rating = ...)
You are not responsible for FYI material in 390.

Just FYI, here is another example of the use of histograms, showing that attending lectures leads to higher grades.

Histograms of test grades for non-attending and attending students.

You will learn about the rest of this plot throughout this class.

All of this suggests that attending 390 lectures is associated with higher test grades. This is from only 4 quarters, but the same pattern exists for every quarter! Of course, things may not be causal.
Here are some of the special shapes that you may come across:

- **Bell-shaped**
- **Skewed**
- **Bimodal**

It turns out there are 2 special hists both of which look exponential:

- \( f_{\text{eq}} = e^{-\lambda x} \) (exponential)
- \( f_{\text{eq}} = x^{-\lambda} \) (power-law)

In addition to \( \text{hist}(x) \), it's also useful to look at \( \text{hist}(\log(x)) \). Sometimes, if \( \text{hist}(x) \) is skewed, \( \text{hist}(\log(x)) \) will be bell-shaped. Usually, looking at \( \text{hist} \) of \( \log(x) \), or \( \sqrt{x} \), or some other transformation of your data, will make the hist more bell-shaped, and that's a good thing, because of easier interpretation and easier Math (later).
When you get an exponential looking hist, the way to determine whether you have an exponential or a power-law histogram is to transform:

If you get \( \frac{\text{freq}}{x} \), then look for \( \frac{\log(\text{freq})}{x} \) in

1. \( \log(\text{freq}) = \alpha - \beta x \Rightarrow \text{freq} = e^{\alpha} e^{-\beta x} = \text{(constant)} e^{-\beta x} \)
   
   ie. The frequency hist. is really exponential.
   
   As a result, the freq. hist is called exponential. (More later)

   In short, \( \text{freq}(x) \sim e^{-\lambda x} \Leftrightarrow \log(\text{freq}(x)) \sim -\lambda x \)

2. \( \frac{\log(\text{freq})}{\log(x)} \), then

   \( \frac{\log(\text{freq})}{\log(x)} = \alpha - \beta \log(x) \Rightarrow \text{freq} = e^{\alpha} e^{-\beta \log(x)} = e^{\alpha - \beta \log(x)} \)

   These hists are said to follow a "power-law". E.g.

   - \( x = \) magnitude of earthquakes
   - \( x = \) population of cities, on the planet
   - \( x = \) length of words, in a book
   - \( x = \) casualties of wars, for different wars

\[ \text{hist}(\log(x)) \neq \text{plot}\left(\frac{\log(\text{hist}(x)\text{ (sums)})}{\text{hist}(x)\text{ (count)})}\right) \]

Because \( \text{hist}(\log(x)) \) gives \( \text{freq} (\log(x)) \) not \( \log(\text{freq}(x)) \).

(See hint)
For each of the following shapes, come up with at least 1 example of a quantity \( x \) (a random variable) whose histogram you expect to be approximately:

a) Bell-shaped (symmetric)

b) skewed (one way or the other)

c) exponential-looking

d) Bimodal

Describe the quantity clearly, and explain in words why you expect the particular shape. If you have data to support your expectation, then go ahead and show the histogram. (For this problem, \( x \) may be continuous or discrete.)

**hw. lect 3-2**: In this lecture there are many examples of random variables that when considered as quantitative, have an exponential-looking histogram. Identify one of them, and plot its rel. freq. histogram (the rel. frequencies are in that lecture, too). By hand