Data organization

Suppose we go out and collect some data. What do we do with it? First we need to figure out what *kind* of data we have. To illustrate, let's do a simple experiment and collect the height and sex of several students in the class. Now we have some data. What kinds of data are there?

Categorical - we can divide this into two types:

1. Not ordinal: you can't sort the categories. Examples might be:

   Sex, blood type, geographic location, color.

   None of these can be put into any obvious order (e.g., do males or females come first? Which blood type comes *first*?

2. Ordinal - even though these are categorical, we can put the data into some type of logical order:

   Size of t-shirts: small, medium, large, extra-large, etc.

   Teacher evaluations: 1, 2, 3, 4, 5. If someone gets two "2's", does that mean we could add them up to get a 4?? Of course not!

Quantitative - here we also have two (related) types:

1. Discrete - technically this means you can list all possible values. Usually (but not always!) we mean integer data. For example:

   Number of eggs laid by a duck.

   Number of bacterial colonies on an agar plate.

   Age (is age really disrete?).

2. Continuous - we can't list all possible values (strictly speaking, this depends on deep concepts involving the number line, but let's keep it simple). Some examples:

   Weight of a person (if measured precisely enough).

   Circumference of a femur.

   Length of a snake.

   Notice that sometimes continuous data are only as precise as the measuring device used. For example, practically, we can't measure weight to infinite precision, but we could imagine that we could measure this to any precision we want if we had the equipment.

So what kind of data did we collect?

Now that we know a little about the types of data, what else can we say about our data? Probably the next most important thing is to talk about sample size. How many *records* do we have?

In our little example, we have a given number of people (our sample size). For each person we have two observations (sex and height). Each person is a record.

Some people (or texts) will use the words *case*, or more rarely *observational unit* for record. We'll stick with record.

Sample size is generally indicated by the letter n. So if we go out and measure 23 snakes, our sample size would be $n = 23$.

So we know about types of data and records. The next most important thing is to take a look at our data. What do they look like? How many tall people are there? Short people? Are most people near the average height?

As statisticians, this is crucial because it lets us decide how to proceed with our analysis. If we wanted to know, for instance, the third tallest person, this would be difficult if we just had a bunch of numbers.

To take a first look at their data, statisticians often sort them. A good way to do this is with a stem and leaf plot. Let's demonstrate by using some (simulated) data on the heart rates of 19 mice:

592 585 608 599 635 694 591 559 628 721 538 651 718 505 635 633 558 586 579

First we sort these data:

505 538 558 559 579 585 586 591 592 599 608 628 633 635 635 651 694 718 721

(Sometimes we can sort and plot at the same time - though usually we let the computer do all of this).

Now we arrange the data into our stem and leaf plot:

50 | 5
51 |
52 |
53 | 8
54 |
55 | 89
56 |
57 | 9
58 | 56
59 | 129
60 | 8
61 |
62 | 8
63 | 355
64 |
65 | 1
66 |
67 |
68 |
69 | 4
70 |
71 | 8
72 | 1

Key: 50 | 5 = 505 bpm

This particular stem and leaf plot is a bit long in that it covers most of the page and has lots of 0 values (values for which there is nothing after the |). It probably makes sense to re-arrange the plot as follows:

5 | 146689999
6 | 01334459
7 | 22

Key: 5 | 1 = 510 bpm

In the second example the data are rounded to the nearest 10 beats (e.g, 579 $\rightarrow$ 580). Both of the above are valid stem and leaf plots, although the second is probably a bit better here. The examples do illustrate that giving a key is important.

Let's do one more example using the same data:

5 | 14	
5 | 6689999	
6 | 013344	Key: 5 | 1 = 510 bpm
6 | 59	
7 | 22	

From this you can see that we can also group our data in groups of 5 (instead of 10). This is perfectly valid; you can even use other groupings (like 2) if you really had a good reason.

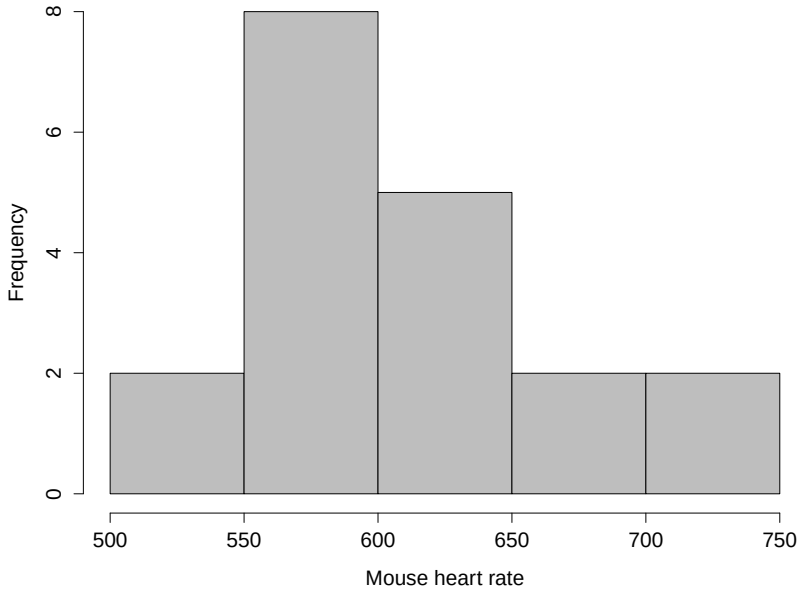
Histograms and barplots:
Histograms and barplots are more graphical methods to look at our data. While a stem and leaf plot is a good initial start, sometimes we want higher quality graphs. They can reveal even more information about our data. Let's deal with histograms first.

Histograms are used a lot and are very useful as they easily lead into frequency distributions (more on distributions later). Histograms are generally only used with quantitative and continuous data.

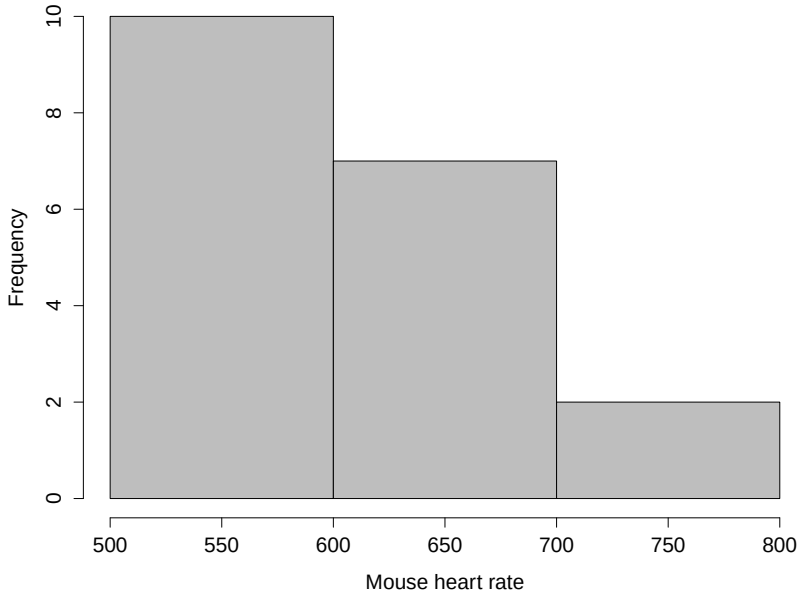
1. Divide your data into groups (similar to what you did above for a stem and leaf plot).
2. Count the number of data points in each group.
3. Draw a x axis that includes the range of your data.
4. Draw a y axis that lists the frequencies for each of your groups (make sure it's high enough to include the category with the highest frequency).
5. Draw bars over each group of the correct height (=frequency).

Let's try making a histogram of our height data.

Here's a histogram of our artificial mouse heart rate data:



Now notice that just like with stem and leaf plots, we can re-arrange our groups and get different looking histograms:



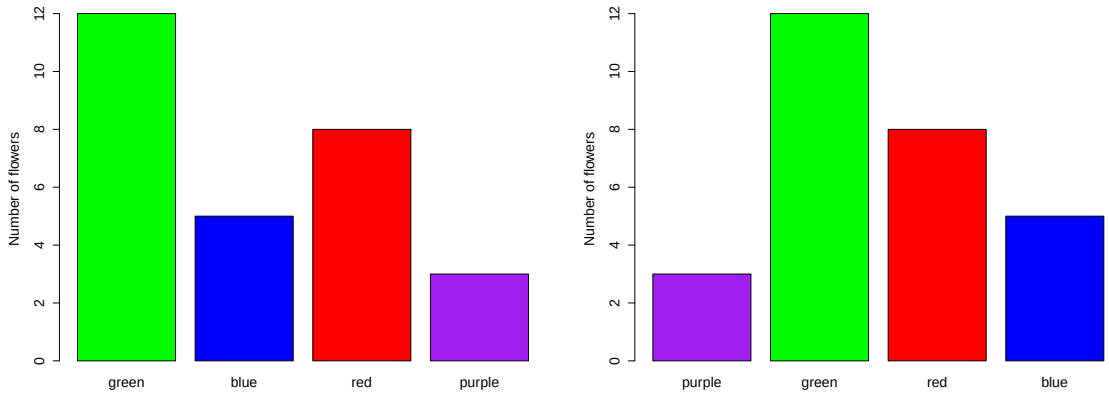
How do we decide how may "groups" we want? There are actually formulas out there that will give us a supposed *optimal* number of groups, but we'll keep it simple. Somewhere between 5 and 15 is probably good, although most of the time we'll let the R figure it out

for us since R has the formulas built in.

Now let's deal with barplots. Barplots are used when the data are not continuous. The advantage of barplots is that they can be used with categorical data as well as discrete data.

(Let's plot our data on sex as a barplot.)

Sometimes it's pretty obvious what order the x axis goes in for a barplot. Ordinal or discrete data have an order. On the other hand, sometimes it's completely arbitrary. Here are some (made-up) data on flower colors:



The barplots are actually identical, but they look different because the order along the x axis is different.

Finally notice that in barplots, the bars should not touch each other (a histogram, which represents continuous data, shows the bars as touching).

Let's summarize the differences between histograms and barplots:

> Histograms are used for continuous data.
>
> Barplots are used for discrete or categorical data.
>
> In a histogram, the bars touch each other to indicate the data are continuous.
>
> In a barplot, the bars do not touch.

Finally, let's use our new found knowledge of histograms (and barplots) to start describing what the *distribution* of our data is.

This is actually quite important, as much time is spent by statisticians in figuring out how data are distributed. This can make a big difference in how data are eventually analyzed.

A distribution describes what our data look like. For example, it tells us How many short people are there? How many tall people? How many "average" people (note that we haven't defined *average* yet?

We visualize this using a graph such as a histogram, and the overlay this with a smooth line. For example:

Let's define a few parts of a typical distribution:

Mode: The highest value of a distribution. Which value (or set of values) has the highest frequency (what's the most common value)?

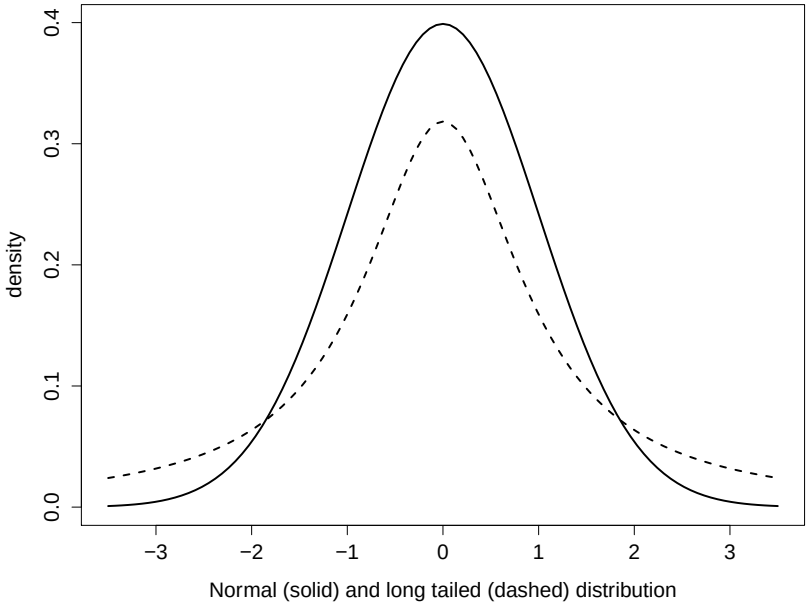
Tails: The ends of the distribution. Usually these are quite skinny (i.e., close to the x axis), but they don't have to be. We'll see some examples of fat tails below.

So what kinds of shapes can we get? The first two are illustrated on the next page, the rest after that.

Symmetric, bell shaped: Very common in biological data. We often call this a *normal* distribution, though we're not ready to explain all the details about this type of distribution for a while. There's also an example on the class web page (see the yellow curve).

This type of distribution is also very important in statistics - we'll explain why later in the course.

Symmetric, with long tails: Although this looks a lot like the previous distribution, this type of distribution can create real headaches for statisticians. See the red curve on the class web page for an example. It is actually possible to tell the difference between this and the above, but the way to do this will have to wait until later in the semester.



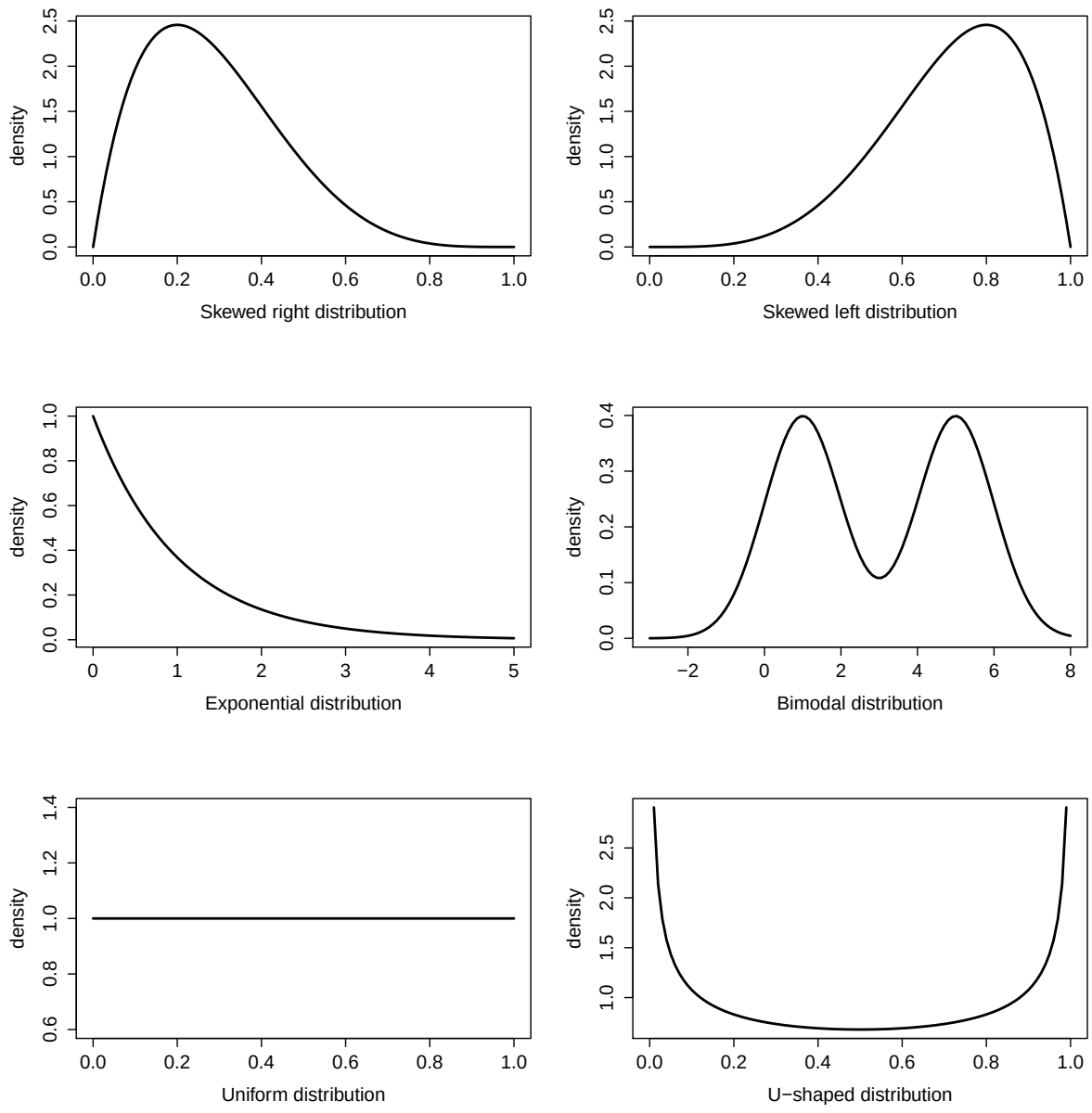
Skewed left (or right): The distribution is no longer symmetrical. A right tailed distribution describes a distribution with a long tail on the right.

Exponential: Actually does have two tails, but the tail on one side is very fat (in fact, it rises up towards infinity).

Bimodal: A distribution with two obvious peaks. Usually the peaks are similar in height, but they don't have to be (if one is higher, it is still called the mode).

Uniform: Every outcome is equally likely - the distribution is simply a horizontal line at a specific height over the x axis.

Symmetric, U-shaped: Both tails are fat, the center is close to the x axis. Strangely, this type of distribution is often much easier to deal with than a long tailed distribution (technically, because the tails stop at a certain point; in a long tailed distribution the tails go on to $\pm\infty$)



The last thing we need to do before we can go on and look at numerical ways to describe our data is to discuss notation. Let's start by differentiating between upper and lower case letters:

Y This represents a variable. For instance, this could be birth weight, estrogen levels, or length of a snake. It does *not* say anything about the actual value of birth weight, estrogen level, etc.

Y is considered to be random, since we don't know what the value is.

y Represents an actual value for Y. For instance, 9 pounds, 13 ounces for birth weight, or 45 pg/mL of estrogen, or 1.23 m for a snake.

Often, we want to know the value for a specific individual (or record), in which case we need to index our y using a subscript:

y_i This represents the value in the i^{th} place.

Since this might be just a little confusing, let's do an example. Suppose you go out and collect the lengths (in cm) of 6 worm snakes (again, the data are made up - the numbers are a bit big for worm snakes):

14 12 16 23 18 17

Now if we want to index this we would have:

$$\begin{aligned} y_1 &= 14 \\ y_2 &= 12 \\ y_3 &= 16 \\ y_4 &= 23 \\ y_5 &= 18 \\ y_6 &= 17 \end{aligned}$$

Often in statistics we want to add things up. So often, in fact, that we need a special symbol because writing out 14 + 12 + 16, etc. gets very tedious. So let's use these data and add them up using the mathematical symbol Σ. Let's do the following and then explain it:

$$\sum_{i=1}^{6} y_i$$

This tells us to add up all the numbers in the positions y_1 through y_6. The $i = 1$ at the bottom of the Σ tells us where to start, and the 6 at the top (we could also use n here since we're adding up all the numbers) of the Σ tells us where to stop. In other words, what we have is:

$$\sum_{i=1}^{6} y_i = y_1 + y_2 + y_3 + y_4 + y_5 + y_6 = 14 + 12 + 16 + 23 + 18 + 17 = 100$$

To be complete, you should realize that Σ can be used in a number of different ways. Here are some examples:

$$\sum_{i=1}^{3} y_i = y_1 + y_2 + y_3 = 14 + 12 + 16 = 42$$

Or, a bit more complicated:

$$\sum_{i=1}^{4} i = 1 + 2 + 3 + 4 = 10$$

Or even:

$$\sum_{i=4}^{5} 3 \times i = 27$$

If you want, you can try the following (answers at the bottom of the page):

$$\text{a)} \sum_{i=3}^{5} y_i \qquad \text{b)} \sum_{i=1}^{5} y_{i+1} \qquad \text{c)} \sum_{i=1}^{5} 15y_i \qquad \text{d)} \sum_{i=1}^{6} (15 - y_i)$$

We'll start to use the Σ symbol a lot in the next set of notes.

a) 57 b) 86 c) 1245 d) -10