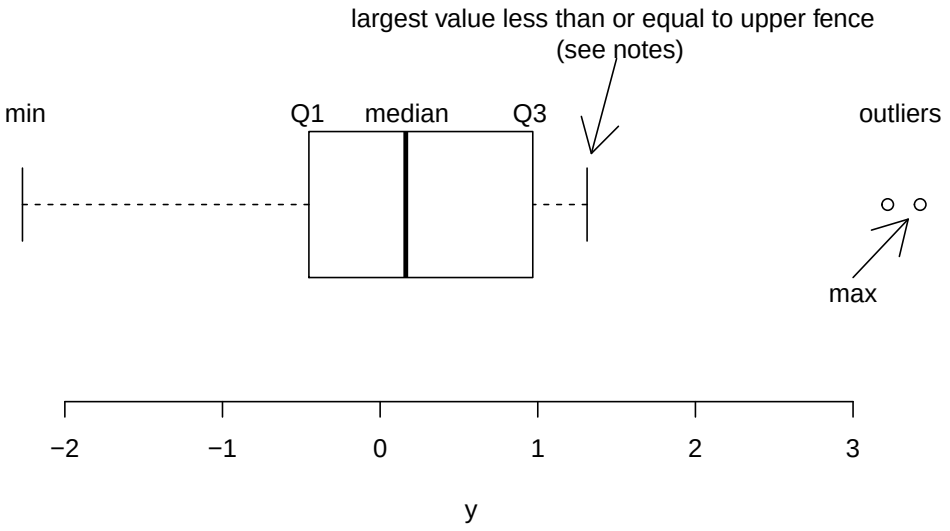
Box plots, populations versus samples, and random sampling

This set of notes covers several important topics that we glossed over previously. We'll start with another type of graph called a boxplot, which we couldn't discuss until we learned about medians.

Boxplots A boxplot is a very nice way of taking a look at our data and figuring out how spread out the data are, where most of the data are, and if there are any outliers.
Here's an example of a boxplot:



Some of these things you should be familiar with:

median, min, max, outliers.

What about the rest (more details below)?

Q_1 = median of the lower half of the data (not including the median).

Q_3 = median of the upper half of the data (not including the median).

(note: another name for the median is Q_2)

Upper fence = $Q_3 + 1.5IQR$

Lower fence = $Q_1 - 1.5IQR$

where $IQR = Q_3 - Q_1$

That's the basic outline, but let's go through this again and provide some details:

1) determine the median.
2) determine Q_1 and Q_3.
    a) divide your data into two halves, upper and lower. Do ***not*** include the median when you do this (just *eliminate* the median from your data when you do this).
        i) if you have an odd number of observations, just leave out the median.
        ii) if you have an even number of observations, the median isn't part of your data, so you can just ignore it.
    b) Q_1 = median of the lower half of your data (not including the median).
    c) Q_3 = median of the upper half of your data (not including the median).
3) calculate the IQR: $IQR = Q_3 - Q_1$.
4) now calculate the upper and lower fences:
    a) get $1.5 \times IQR$.
    b) lower fence = $lf = Q_1 - 1.5IQR$.
    c) upper fence = $uf = Q_3 + 1.5IQR$.
5) now draw the actual boxplot:
    a) draw a horizontal (or vertical) line (an axis) going from somewhere below the minimum value to somewhere above the maximum value.
    b) don't forget to add tick marks and actual y-values on the axis.
    c) draw lines (perpendicular to the axis) for the median, Q_1 and Q_3.
    d) draw a box extending from Q_1 to Q_3.
    e) draw a line (parallel to the axis) going to the minimum data value that is $\geq lf$. Then add a short perpendicular line at the end of this line.
    f) draw a line (parallel to the axis) going to the maximum data point that is $\leq uf$. Again, add a short perpendicular line at the end of this line.

        Important do NOT draw the fences on your plot.
    g) any values that are outside the fences (values bigger than the upper fence or smaller than the lower fence) are outliers.

        draw individual dots for any outliers.

This is not the way most people make boxplots, but it is pretty close and the calculations are much easier. If your sample size is reasonable, the differences in the resulting plot are pretty minor.

©2016 Arndt F. Laemmerzahl

Optional: what most people (and R) do is to not exclude the median. In this case, the median counts 1/2 of a data point. For example, if we have 7 data points, the median is the fourth data value. To calculate Q_1, we now assume we have 3.5 data points in the lower half, and get then divide this in half to get 1.75. In other words, Q_1 corresponds to the 1.75^{th} data point (which we then have to calculate). As you can tell, the math gets rather more complicated, which is why we're taking a simpler approach here.

Before we get confused, let's do an example. Here's some (real) data on the length of radish seedlings exposed to caffeine:

14 5 13 10 12 16 6 24 13 33 16 12.5 13.5 1.5 15.5 30

The first step is to sort the data:

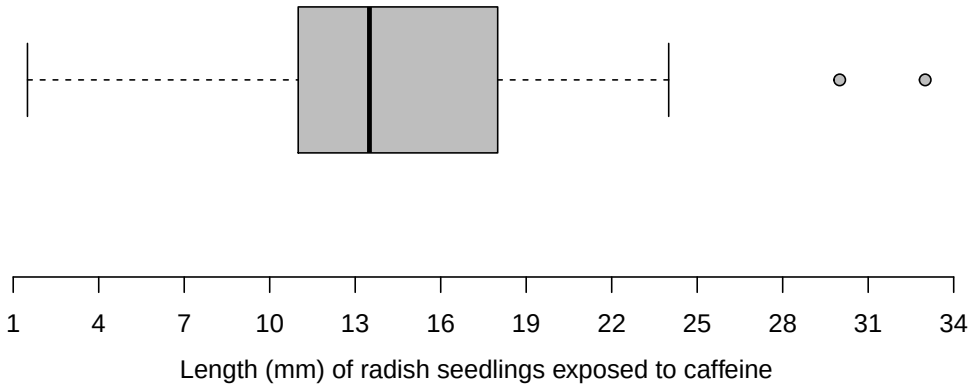
1.5 5 6 10 12 12.5 13 13 **13.5** 14 15.5 16 16 20 24 30 33

And we find the median is 13.5. We then get $Q_1 = (10+12)/2 = 11$ and $Q3 = (16+20)/2 = 18$ (we do not include the median in these calculations).

Then we get: $IQR = 18 - 11 = 7$ and $1.5 \times IQR = 1.5 \times 7 = 10.5$. This gives us the fences: $lf = 11 - 10.5 = 0.5$ and $uf = 18 + 10.5 = 28.5$.

We notice that the minimum = 1.5 which is greater than the lf, so we have no outliers on the low end.

On the other hand, we notice that there are two values that are greater than the uf (30 and 33), so we have two outliers on the upper end. Let's draw the actual boxplot:



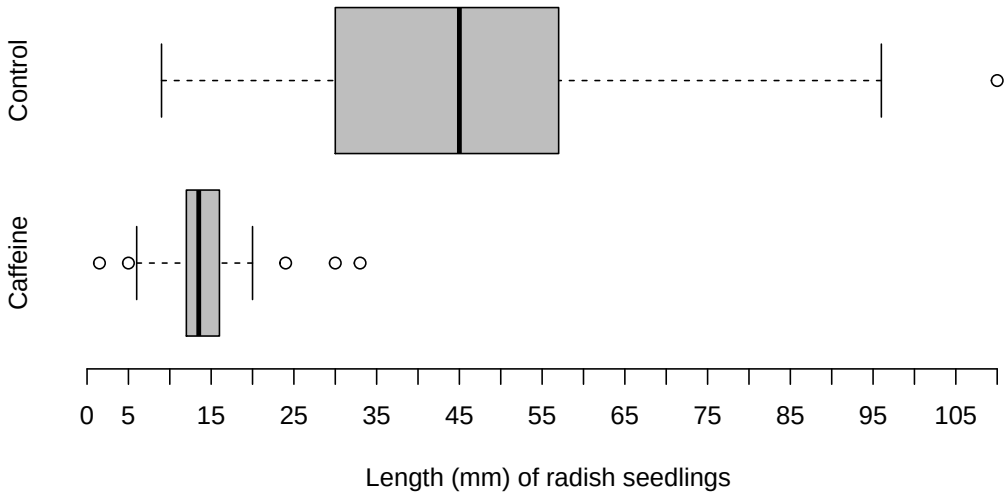
©2016 Arndt F. Laemmerzahl

This plot is drawn using our method of doing boxplots. It is obviously possible to get R to draw boxplots our way, but it isn't easy. From here on, we'll let R do things the way it wants to. You'll see in the example below that our radish boxplot looks a little different with the default method in R. (Notice the boxplot labeled *Caffeine*. If you ignore the differences in scale, you will see that the R version shows more outliers than our version).

Before we go on, we should point out that sometimes other information is added to a boxplot as well.

> Sometimes a dot is used in the middle of the box to indicate the mean.
>
> Confidence intervals (more on these later) can be added.
>
> Other variations exist as well. Hopefully the person doing the boxplot will provide a key if they are adding extra information.

Boxplots can also be used to compare different populations. For example, what can we say about radish seedlings that were not exposed to caffeine? Let's do a *parallel* boxplot and compare our caffeine exposed radishes to a group of control radishes:



Notice a parallel boxplot uses the *same* axis for both plots. Looking at this plot, it does look like there's an obvious difference in length (and variability!) between the seedlings exposed to caffeine and the control group. To make sure, though, we'll need to back that up with statistical procedures that will be covered later.

The boxplot for the caffeine exposed seedlings in this parallel boxplot uses exactly the

©2016 Arndt F. Laemmerzahl

same data as we used before. In this case the boxplot was generated by the default method in R and actually does look rather different (ignore the scale, but notice all the extra outliers). This is a bit unusual as most of the time the default method in R and the method we learned above will only have minor differences.

Samples and populations. Now that we have boxplots figured out, we need to discuss one of the most important ideas in statistics - that of samples and populations. Let's start with an example:

> *Example*: Suppose we tried to figure out the weights of everyone on campus. George Mason University has a student body size of about 34,000. How could we do this? Is it even possible to weigh everyone in a reasonable amount of time?
>
> It should be obvious that if this isn't impossible, it's at least very difficult.

Let's try another example:

> Let's try to get the weight of all rabbits in Northern Virginia.
>
> Again, it ought to be obvious that this is impossible.

One more example:

> Count every word in the statistics notes posted on line.
>
> This might be possible, but it's certainly not practical. Besides, it's highly doubtful you'd come up with the exact right answer just by counting all the words.

We can't really measure the entire population in any of the examples above. Instead, we decide that we'll take a *sample* from this population and use it toe *estimate* what we're interested in. Here's how we could deal with the above problems:

> We pick 100 people at random from GMU and weigh them.
>
> We catch 30 rabbits at random locations throughout Northern Virginia and weigh them.
>
> We pick 20 pages at random from the notes and count the words.

Once we have our sample, we can then use this to estimate the things were interested in in our population. This type of estimation is sometimes called statistical inference - we make conclusions about a population based on a sample. Sampling actually has several advantages over trying to measure the entire population:

> It's often much quicker.
>
> Sampling is often cheaper.

©2016 Arndt F. Laemmerzahl

> Although it might seem contradictory, sampling is sometimes more accurate (think of the example of counting words in the notes).
>
> Often it's impossible to measure the entire population (e.g, the rabbit example).

Surprisingly, the whole issue of samples vs. populations has even been addressed by the Supreme Court as it applies to the 10 year census. We'll give some details on this in class. In the meantime, let's talk a bit more about populations.

We need to define the population we're interested in carefully. Suppose we were trying to figure out the color of people's hair. What is our population? GMU? The United States? Asia? The World?

We should try to be a bit more specific; since we just said people's hair color, the best answer is World, but that's probably not what we're really interested in.

Once we've decided on what our population is, we need to make sure we take our sample from *this population*. If we're trying to figure out hair color of people in Norway we can't sample people in Asia!

Here's a good example of sampling done correctly:

> Suppose we're interested in the number of people 6 feet or taller in New York City. We randomly select 250 people that live in New York, measure their height, and write down whether or not they are taller than 6 feet. Our population is all the people living in New York (it'd be impossible to measure everyone). Our sample is selected from this population.

But sometimes our sample is a bit weird, and it's not easy to see what our population is.

> Let's feed gerbils to cats. Suppose we were interested in trying (for whatever reason) to figure out how many gerbils a cat can eat at one time. We take 15 hungry cats that we have starved for 48 hours and start feeding them gerbils in a lab. The number of gerbils each cat eats within an hour is written down.
>
> What is the population?
>
> The population is the number of gerbils a cat can eat under conditions similar to this experiment. The population here is not naturally occurring, but rather something we set up in a lab. How can we apply this information to the real world? This is potentially difficult and needs to be done carefully!

Although most lab or zoo studies are done carefully and are interpreted in a reasonable way, one should still be aware that sometimes it's note clear how the samples relate to the population.

©2016 Arndt F. Laemmerzahl

Example: saccharin, the artificial sweetener, used to have cancer warnings on the packages. Why did these warnings disappear? Because the warnings were based on a sample taken from rats, which does not apply to the population (humans). Rats don't react the same way to saccharin as humans do.

Still, as mentioned, a tremendous amount of vital and important research is done in labs or zoos!

Estimates and parameters. This whole idea of samples and populations does lead to some important distinctions. We can't measure the weight of everyone on campus, or weigh all the rabbits in Northern Virginia. The result of this is that we don't actually know:

the true average weight of everyone on campus.

the true average weight of rabbits in Northern Virginia.

All we can do is use our sample to *estimate* these quantities. However, these quantities (the true average weight of everyone on campus, for instance) do exist. We just know know what they are.

Even if we don't know what they are, we still need to refer to these quantities. After all, we're really not interested in the *sample* average, we're interested in the *population* average. We don't want to know the average weight of people in our sample, we want to know the average weight of everyone on campus. So we refer to our population *parameters* as follows:

The true average (or true standard deviation), and so on, are termed population parameters:

The true population mean (which we dont know) is symbolized with the Greek letter:

mu or μ.

Similarly, the true population standard deviation is symbolized by the Greek letter:

sigma or σ.

Since we don't know what these are, we estimate them with the sample mean and sample standard deviation:

$\bar{y}$ estimates μ.

s estimates σ.

One obvious question at this point might be might be "how good are our estimates?" We'll be able to figure this out, but it will have to wait for a later date.

©2016 Arndt F. Laemmerzahl

At this point you might be tempted to think that population parameters are symbolized by Greek letters and estimates by Latin letters. While this is often true, it doesn't always work out that way. Let's take a look at proportions, for instance.

> Let's try to figure out the proportion of people infected with HIV. If we use the Greek letter for p as the population proportion, we would need to use π. As you can guess, this isn't going to work, since π is already taken by 3.1415926... (some statisticians don't care about this and will use π anyway, which can be confusing in an introductory class). Since we can't use π, we will use p instead. That's the Latin letter.
>
> So what is the estimate? The estimate is given by $\hat{p}$ (we call this p-hat).
>
> In other words, for proportions we have:
>
> > $\hat{p}$ estimates p.
>
> Incidentally, the ˆ (hat) symbol always means "estimate" in statistics, so anytime you see it used over a variable (even a Greek letter!) you know it's an estimate.

Random sampling. The last thing we need to do is to figure out how to take our sample in such a way so it truly represents the population. The best way (the gold standard in statistics) is to use random sampling. This isn't always possible, but it really is the best. Let's discuss random sampling a bit and then talk about some other sampling techniques.

Random sampling takes place in such a way so that every item in the population has exactly the same probability of being in the sample as any other item. In addition, picking any one item in the population should not influence the probability of picking any other item in the population.

One possible way to do this is to assign every item in our population a number and then pick n random numbers until we have filled our sample. Let's illustrate this with our population of GMU students.

> First, we number all 34,000 students. We could simply get an alphabetical list, and then assign the students the numbers 1 through 34,000 in order.
>
> If we want a sample of size $n = 100$, then we pick 100 random numbers.
>
> One way of doing this (a bit old-fashioned) is to use a random number table.
>
> > We go into a random number table and pull out 5 digit numbers.

©2016 Arndt F. Laemmerzahl

Since the numbers are random we can start in any row or column, go in any direction we like, etc. It makes no difference.

The only rule is that we probably shouldn't use the same number(s) twice.

Any number that's bigger than 34,000, we simply discard.

We keep going until we have 100 random numbers between 1 and 34,000 (inclusive).

For really small samples (say $n = 10$ or less, this is still a reasonable method. These days it's much easier to get random numbers from a computer. Here's how to sample 16 numbers from a population of 150 items in R:

```
# Step 1:  list all the items in R
# (for example, this puts all the numbers from 1 to 150 into y
y <- (1:150)
# (optional:  type y at the command prompt if you want to
# see your sequence of 150 numbers)

# Step 2:  take a sample of size 16:
# takes a sample of size 16 from y
y2 <- sample(y,16)
# (optional:  type y2 if you want to see your sample)

# Step 3:  if you want to make it even easier:
sort(y2)
```

As it turns out, computer generated random numbers are actually *pseudo-random*; there are whole books written on this topic. A computer can not generate a sequence of totally random numbers. A simple example to see this is to set the *seed* in R before asking it for random numbers.

In the above example, add the following line at the top:

```
set.seed(2345)
```

If you do, and you run the code in the example over and over, you'll see that you get exactly the same sequence of numbers.

Normally, R tries to give you a different set of random numbers every time by using something like the computer time as a seed. However, technically this still isn't random.

As an aside, until recently the random number generator in Excel was one of the worst ones we knew about (the random numbers generated by Excel show patterns and other problems). Microsoft has changed its random number generator recently, but no one seems

©2016 Arndt F. Laemmerzahl

to know exactly what's going on now.

So let's summarize what we learned about sampling and add a few other bits:

1. You should make sure you do correct random sampling. Define your population carefully, then make sure you sample from this population.

2. Make sure you do random sampling if at all possible (see below).

3. Generally, the larger your sample size, the better. We'll explore this topic further.

4. Sometimes it simply isn't possible to do random sampling. Statistical sampling is a special topic in statistics - whole books are written on this. Here are two common exceptions to pure random sampling:

    a. *Systematic sampling.* In this case we sample every n items instead of taking a random sample. For example, we sample trees every 25 meters. This can be a huge time saver - and in any case we can't really number every tree in a forest).

    b. *Opportunistic sampling.* Here one samples what one finds. If you're investigating endangered turtles you probably want to include every turtle you find in your sample. After all, they're probably rare and you need to take advantage of every turtle you find.

    Opportunistic sampling does have some issues (it's not random), but often it's all you can do.

    Most sampling methods do have some kind of random component to them, although it's sometimes hard to see.

©2016 Arndt F. Laemmerzahl