**Miscellaneous topics & random sampling**

Before we go on, we need to pick up some pieces.

I. Boxplots (see section 2.5 *{2.4}*in text (but note that we will *always* make the "modified" boxplots)).

- Boxplots are a way of summarizing the data in a very quick and understandable fashion.

- We needed to understand medians before we can do boxplots.

- Problem: boxplots can be done differently by different people. Your text presents one way (probably the most common), but there are others.

- The basic idea is as follows:

1) determine the median.

2) determine $Q_1$ and $Q_3$.

- what are $Q_1$ and $Q_3$?

- the median divides the data in half.

- $Q_1$ divides the lower half in half, $Q_3$ divides the upper half in half.

- we do NOT include the median when figuring out $Q_1$ and $Q_3$

- incidentally, the median is also known as $Q_2$.

3) multiply the IQR (interquartile range, = $Q_3$ - $Q_1$) by 1.5.

4) Draw a horizontal (or vertical) line (an axis) going from somewhere below the minimum value to somewhere above the maximum value.

5) Draw a line for the median.

6) Draw a box extending from Q1 to Q3.

7) Now draw a very light mark (it shouldn't be visible in the final plot) at $Q_1$ - 1.5 IQR. This is called the "lower fence".

8) Draw another light mark (again, it shouldn't be visible in the final plot) at $Q_3$ + 1.5 IQR. This is the "upper fence".

9) Draw a line going to the minimum value that is bigger (greater) than the lower fence. This may or may not be the smallest value in your data.

10) Draw another line going to the maximum value that is smaller (less) than the

upper fence. Again, this may or may not be the largest value in your data.

11) If any values are outside the upper or lower fences (e.g., you have values bigger than the upper fence or smaller than the lower fence) draw individual dots for these.

- Best to look at an example:

See example 2.21 on p. 40 **[2.21, p. 33]** *{2.4.2, p. 46}* of your text.

Here is exercise 2.24 from p. 47 **[2.31, p. 38]** *{2.4.2, p. 51}*. MAO is an enzyme that may be involved in schizophrenia, and these levels were measured (in nmoles/$10^8$ platelets) in 18 patients:

1) Determine the median:

- arranging the data from largest to smallest:

4.1
5.2
6.8
7.3
7.4 -
7.8
7.8
8.4
8.7 _
9.7
9.9
10.6
10.7
11.9 -
12.7
14.2
14.5
18.8

- the median is (8.7+9.7)/2 = 9.2

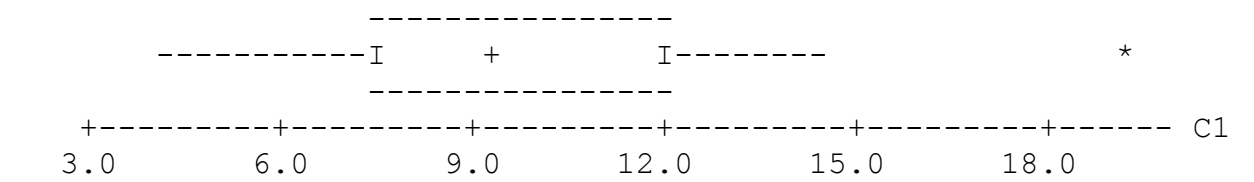
- Q1 is 7.4

- Q3 is 11.9

- the IQR (interquartile range) is Q3-Q1 = 4.5

- multiply this by 1.5: 4.5 x 1.5 = 6.75

Now arrange all this into a boxplot (we'll go over the details in lecture):




As mentioned, not all boxplots are alike:

- sometimes the mean is used

- sometimes CI's (confidence intervals) are used instead of IQR's for the length of the box.

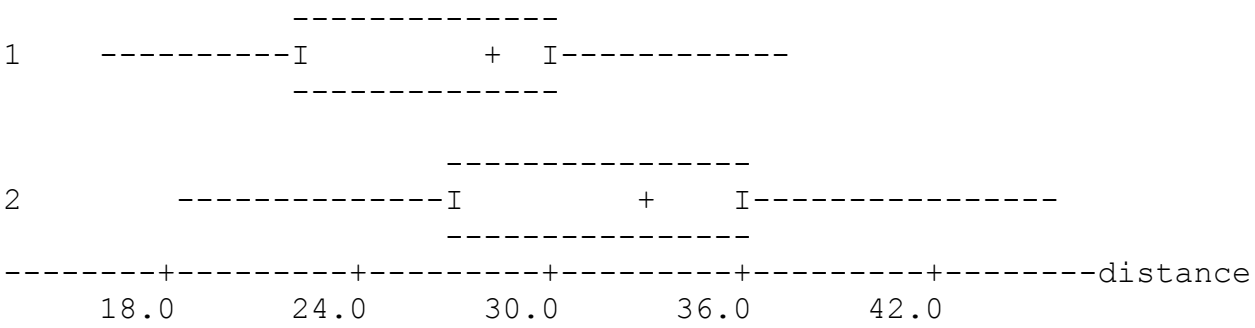
- sometimes the whiskers go to the extremes, even if these are outside the range we use above.

- so, hopefully the person who made the boxplot will include a key to tell you what all the things are.

- to make things worse, there are slightly different ways of getting $Q_1$ and $Q_3$. The differences are minor, but can lead to some confusion (one reason we're not using Minitab anymore is because it does it two different ways (and doesn't tell you!!).

- R does boxplots close to this description, but it is a bit different (it calculates $Q_1$ and $Q_3$ a little differently). At least it only does them one way unless you tell it to do something different.

Boxplots can also give you a good initial comparison between two different populations. See p. 46 **[35]** in your text. Also, here's part of exercise 7.84 **[7.102]** *{7.S.18}*. In this example, researches wanted to know if infection with malaria hurt the ability of lizards to run. They measured how far (meters) the lizards could run in two minutes:




(1 is infected, 2 is uninfected)

One can tell almost right away that there is probably an impact of malaria on lizards.

II. Samples and populations.

Suppose we tried to figure out the weights of everyone on campus. How could we do this?

1) Weigh everyone. Is this practical? Possible? Accurate?

- try counting every word in your textbook. You think you might do better by estimating?

- a recent (?) debate in congress and the supreme court dealt with this issue as it applied to the census.

2) Pick some people at random (more on this soon) and hope that they represent, in some way, the people on campus.

- figure out what you want based on this sample.

- advantages:

- easier, less expensive, possibly more accurate.

Often it is impossible to "measure" a whole population, even if we wanted to.

So, we ESTIMATE what we want about our population. This is sometimes also called *statistical inference* - making conclusions about a population based on a sample.

A. Populations:

We need to define this carefully. Suppose we were trying to figure out peoples' hair color. What is our population?

- GMU?

- United States?

- Asia?

- Europe?

- World?

Once we know, we can figure out how to get a sample. Trying to figure out hair color of folks up in Norway isn't going to work if we sample Asia.

Some examples:

- book lists blood types in England. Sampling 3,696 people in England is probably a good way of trying to figure out the overall proportion of blood types. Our population is the proportion of blood types in England.

- feeding gerbils to cats. Suppose someone was interested in trying (for whatever reason) to figure out how many gerbils a cat can eat at one time. The person takes 15 hungry cats, and starts 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.

  - Note that the population here is not "naturally occurring", but rather something set up in a lab.

- See also examples 2.45 *{2.8.3}* and 2.46 in your book.

In general, most often what we are trying to do is to make conclusions about populations where we can't measure everything. We are not often interested in conclusions about small groups such as the:

- height of people in this class

- weight of elephants at the National Zoo,

- etc.

Usually things like this would be regarded as part of a sample, because what we're really interested in might be:

- height of people on campus

- weight of captive elephants in the United States.

Note that in these cases the "samples" are not really random. We'll discuss what this means soon.

III. Estimates and parameters.

In general, we usually don't know the "true" average or standard deviation of a population.

- Can we really measure the height of everyone on campus?

The "true" population mean (which we don't know) is symbolized with the Greek letter

mu or "$\mu$".

Similarly, the "true" population standard deviation is symbolized by the Greek letter sigma or "$\sigma$".

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

$\bar{y}$ *estimates* $\mu$ (*mu*)

$s$ *estimates* $\sigma$ (*sigma*)

How good are our estimates? Chapter 6 will tell us how to figure that out.

"True" population characteristics are often symbolized by Greek letters and termed "parameters".

Sample characteristics (or "statistics" which we use to estimate parameters) are often symbolized by Latin letters.

But not always:

If we're looking at a proportion, we might consider the proportion of people infected with AIDS, "*p*".

*p* is the true unknown proportion.

we try to estimate *p* with *p*-hat, or $\hat{p}$ .

some people do use $\pi$ (pi) for proportions, but that letter is generally thought to be "taken" by 3.14159.....

The "^" (hat) symbol almost always means "estimate", but a lot of the other stuff varies depending on who writes the book (and even sometimes in the same book!)

so $\hat{p}$ estimates p, the true proportion.

See also p. 70 - 72 **[61-63]** *{76-77}* in text.

IV. Random sampling.

Problem - we need to make sure our sample is representative of the population.

A possibility:

- number every item in the population

- pick n random numbers

- sample those items that match the random numbers

The idea:

- pick the sample in such a way so that each item in the population has the same probability of being picked.

- If I pick item *x*, it does not influence the probability of picking item *y*.

Random numbers:

1) for a small sample, use a random number table.

- try to pick a "random" starting point in the table.

- pick the appropriate number of digits

  - if you have 150 items in your population, you should pick three digit numbers.

- ignore any number that doesn't fit your sample

- if you have 150 items, ignore numbers larger than 150.

- continue until you have however many numbers you need.

  - you can pick the second, third, etc. numbers simply by moving to the left, right, up or down. It's irrelevant.

2) for bigger samples, use computer generated random numbers. R will easily generate random numbers for you (look up the runif command from the command line).

- [an aside - computer generated random numbers are actually "pseudo-random". There are books written on this topic, including one by my major advisor!]

- The random number generator in Excel is one of the worst ones we know about (the numbers don't appear truly random). Microsoft has been told about this, but refuses to fix it.

Other comments.

- you want to make sure you actually do correct random sampling.

  - the population needs to be carefully defined, and then one needs to sample randomly from it.

- A few brief comments on sample size:

    - in general, a larger sample is better.

    - you should have a large enough sample so that inferences (=conclusions) about the population are believable.

    - we'll learn a bit more about this later.

- There are actually other ways of taking a sample (systematic sampling, for example), but we don't have time to go into these.

    - there are whole texts written on sampling.

    - all of them will have some kind of random component, though it may not be apparent.