Imagine a data set from a hospital for post-surgery recovery times in days. Seven patients were randomly divided into a control group of three that received standard care, and a treatment group of four that received a new kind of care. The data is

```
> control <- c(22, 33, 40)
> treatment <- c(19, 22, 25, 26)
```

with summary:

```
> summary
```

```
          mean   sd
Control   31.7 9.07
Treatment 23.0 3.16
```

A natural question is

Is the treatment better than the old treatment (the control)?

How might we answer this? Clearly for this data the average of the treatment is much less (31.7 to 2.03 or a difference of 8.7 days), so for this data it was better on average.

But these 7 patients have recovered, really we'd like to extend this comparison to the entire population of people who would be eligible for this care. Now a natural question is

Can we do this?

Question 0.1. Come up with a few reasons why you might not be able to infer information based just on the 7 patients to the wider population.

Consider Figure 1 which shows for a *comparative experiment* what conclusions can be hoped to be drawn based on the design. Recall, randomization allows us to use the language of probability to infer things about the population.

From the description above of the experiment, we do not know if the initial 7 pepole were randomly selected. Likely this is not true. What we are told is that the 7 people were randomly assigned to the treatment groups. Now we might ask

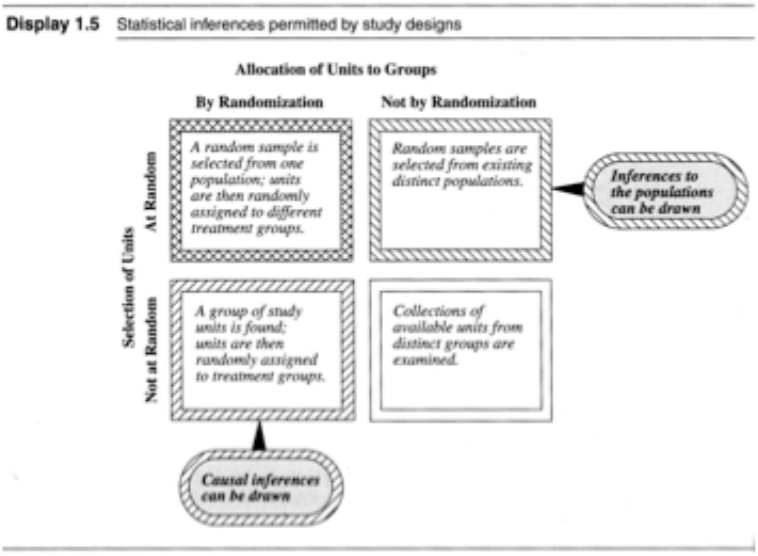


Figure 2: A chart from Ramsey and Shafter (2002).

Figure 1: Graphic showing what can be done through randomization

Is the difference between the groups due to the difference in treatments?

If we were to look at our question: is there a difference, a skeptic might say this:

> There is no difference, you just put the right people into the treatment group and the wrong people into the control group.

Of course, if we know ahead of time, we could have done this. But the random selection is to ensure that *it is unlikely* that this would have happened. However, the skeptic may not believe still, so we will investigate. How? We say:

> Let's assume – for the moment – you are right, there is no difference. Let's consider all possible assignments of subjects to treatment groups and look at how unlikely the difference of 8.7 is.

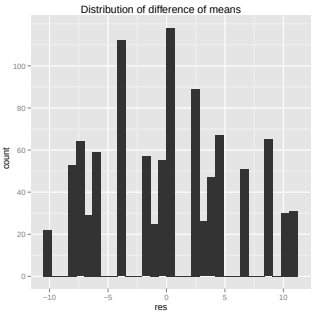
Oh boy, something we can do on the computer. R makes this task manageable. Rather than look at all possible re-arrangements, we take the simulation approach and look at lots os simulations of different randomizations. Here are some commands to do this:

```
> allData <- c(control, treatment)
> SRS <- sample(1:length(allData), 3)
> SRS

[1] 4 2 3

> mean(allData[SRS]) - mean(allData[-SRS])

[1] 6.92
```

These commands take a sample of size 3 (`SRS` which is the 4th, 2nd and 3rd scores) and assign those to the control group and the remaining 4 (`-SRS`) to a treatment group and then looks at the difference of the means. The value of 6.92 says the control was 6.92 days more than the treatment for this rearrangment.

We want to do this 1,000 times say, so we put it into a loop as follows:

```
> res <- c()
> for (i in 1:1000) {
+     SRS <- sample(1:length(allData), 3)
+     res[i] <- mean(allData[SRS]) - mean(allData[-SRS])
+ }
```



Now we have 1,000 such numbers. Their distribution is shown in Figure

We ask, how unlikely is our value of 8.7? Well we could look for when that value comes up exactly, but if the treatment works values to the right of that number (like 10) would also support the statement the treatment works, so we should include those. To see, we have

```
> sum(res >= 8.7)/1000

[1] 0.061
```

So in our simulation 6.1% of the rearrangments produce a bigger difference of control minus treatment. Now we have to ask

> Is there statitical evidence that the difference we saw is large enough given that there is considerable variation due to randomization?

Or something like that. The point is randomization – that which helps us draw conculusions – forces us to think that *perhaps* differences are due to the process of randomizing and not an actual difference in the populations.

Question 0.2. How unlikely does something have to be before you think of it as unusual? For instance, how tall must someone be before you think they are really tall? How expensive a car must it be before you think that is an expensive car? How long is it before you think I'be been waiting in line for a long time? Now try and translate those ideas into how often these unusual things happen relative to all the times they could.

Question 0.3. Historically – way back when – people used to think if something happened as rarely as 1 in $\$10,000$ it was an act of god. Can you think of something that happens that infrequently, but that does not require divine intervention?

In statistics, we give a name to the value 0.061 we computed above. It is called a p-value:

> The p-value answers: how likely are we to see the observed value *or something more extreme* assuming there is no difference.

If p-values are small, then we have evidence that our assumption incorrectly describes the data, if not small we have no reason not to believe the assumption (no difference) describes our data. A generally agreed upon measure of small is $\alpha = 0.05$.

Question 0.4. Since 0.061 is more than 0.05 what can we say about our skeptic's assumption of no difference between the two treatments?

Let's repeat the above with some new data. Suppose we want a MTH 214 professor has 9 students in his class. (Yes a dream scenario for the students.) To make sure there is no copying going on during an exam, he chooses 5 students at random to get form A of an exam and 4 to get form B. For his choice, the data was

```
> formA <- c(50, 60, 55, 70, 65)
> formB <- c(60, 80, 90, 80)
```

The difference between the two is:

```
> mean(formA) - mean(formB)
```

```
[1] -17.5
```

Should the instructor worry that form A was much harder as the average score is more than 17 points less than that on form B?

Question 0.5. Combine the two data sets into a single one. Write down the command.

Question 0.6. Select a random sample of size 5 from the 9 possible indices, following the command above that used `sample`. Call them `SRS`. What are your five indices?

Question 0.7. What are the five values corresponding to your indices? These correspond to a possible form A group, were there no difference in exams and a different randomization was used.

Question 0.8. What is the mean of the five values corresponding to your sample?

Question 0.9. What is the difference in the mean corresponding to your sample?

Question 0.10. Now follow the pattern above to general 1,000 such numbers (differences).

Question 0.11. What percent of your values are less than -17.5? This is the p-value.

Question 0.12. Write a sentence or two in response to the statement

> The two tests are equally hard and should produce identical scores.

What we are discussing here are called *permutation tests*. The text has a web appendix, Chapter 16, that covers this material (start on page 41). This file is on the class website. The book claims that Verizon uses permuation tests to investigate repair times using custom software based on S-Plus (S-Plus is now a pay version of R). So if you finish this project, you can go apply to Verizon! Anyways, lets imagine that we have repair time data for Verizon and company B given by:

```
> verizon <- c(111, 131, 41, 67, 163, 7, 256, 174, 29, 64)
> companyB <- c(52, 119, 209, 209, 166, 25)
> c(mean(verizon), mean(companyB))

[1] 104 130
```

Question 0.13. Verizon claims it has shorter repair times based on this sampe. Is this difference (26 minutes shorter) due to this fact, or is it merely an artifact of sampling variation? Assume a dispatcher randomly assigned either Verizon or companyB to each job.

Sketch out the steps you take to investigate this question. Include the p-value that you find.