Research Methods 2

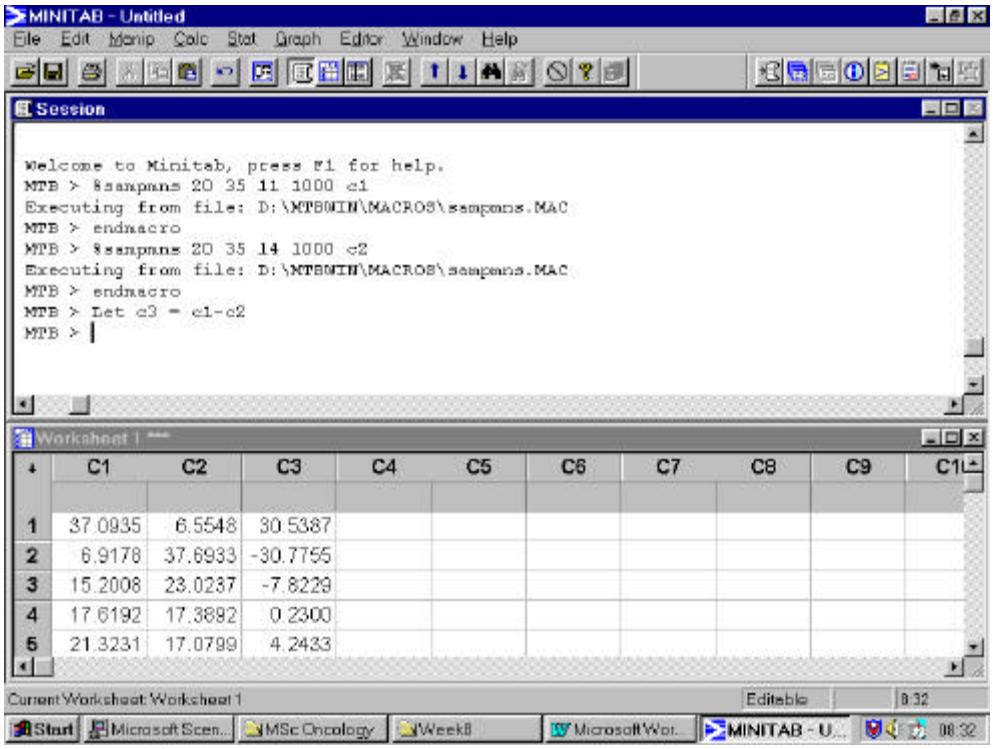
Week 9: Exercise Sheet 1

Solution sheet

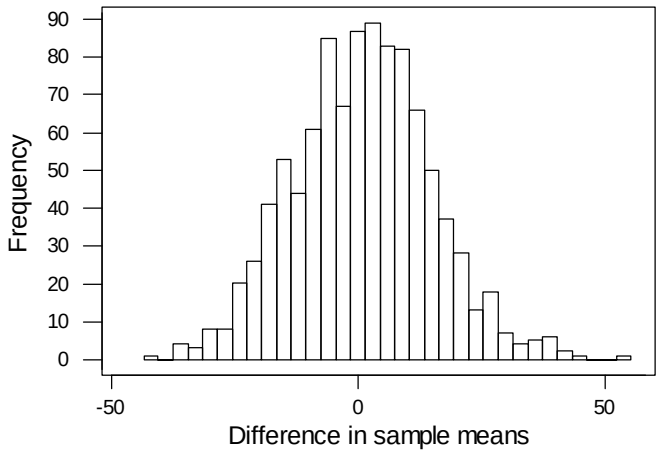
[Again remember that in questions which rely on the random generation of data, the precise numerical values you obtain will differ slightly from those below. However, as usual, the key features of the results should be the same as yours]

Question 1

A typical appearance of the screen after generating the data is show below.

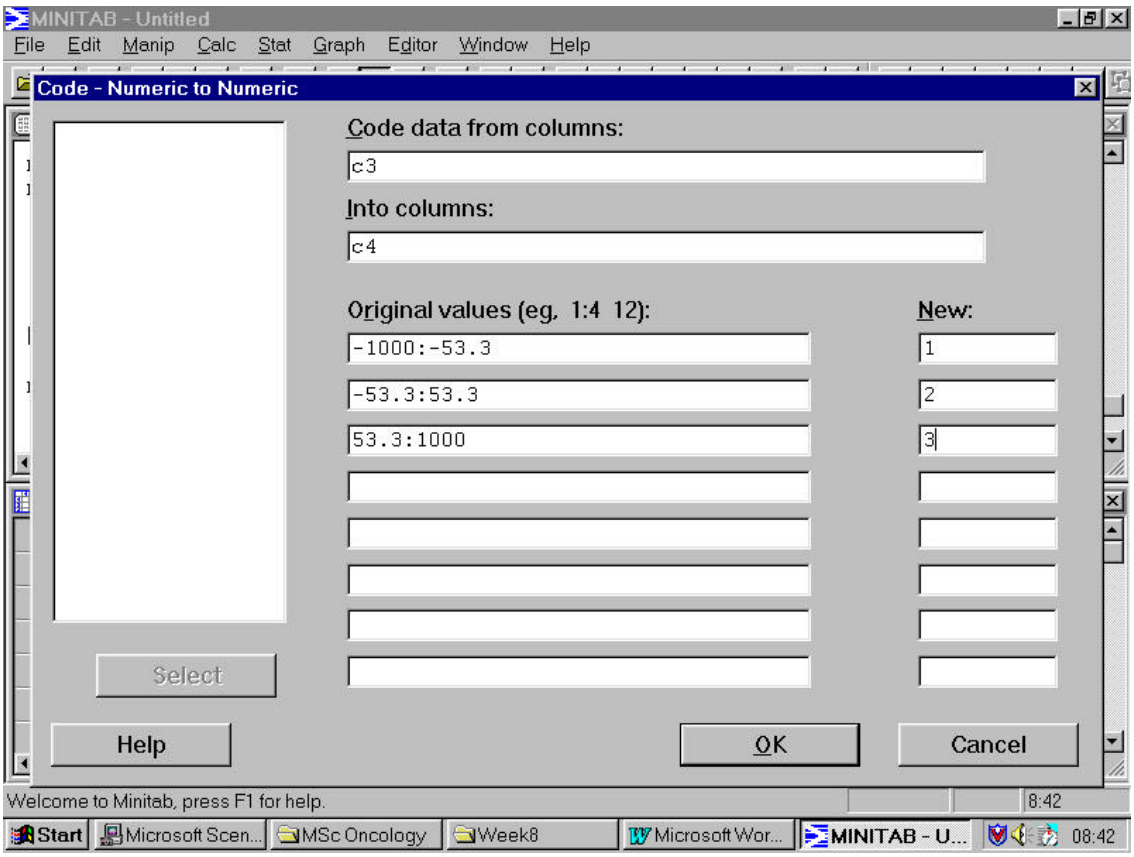


A histogram of the differences stored in C3 is shown below



It is clear from inspection of this graph that a difference of 53.3, which is the difference Takeuchi *et al.* observed, is a very unusual observation if the population means are actually the same. Precisely how unusual requires us to find the numbers of differences in the ranges specified in the question.

Applying the technique described in the hint in the question, you should end up with this screen.



On clicking on **OK** and applying **Stat** -> **Tables** -> **Tally**... we find that all 1000 differences lie between –53.3 and 53.3. In other words a difference as extreme as that we observed has not been seen among 1000 differences that have been generated *under the assumption* that the population means are equal. We do not say that the P-value is 0, since a value more extreme than 53.3 might have been seen if we had, e.g. generated 10000 samples. What we do claim is that $P < 0.001$.

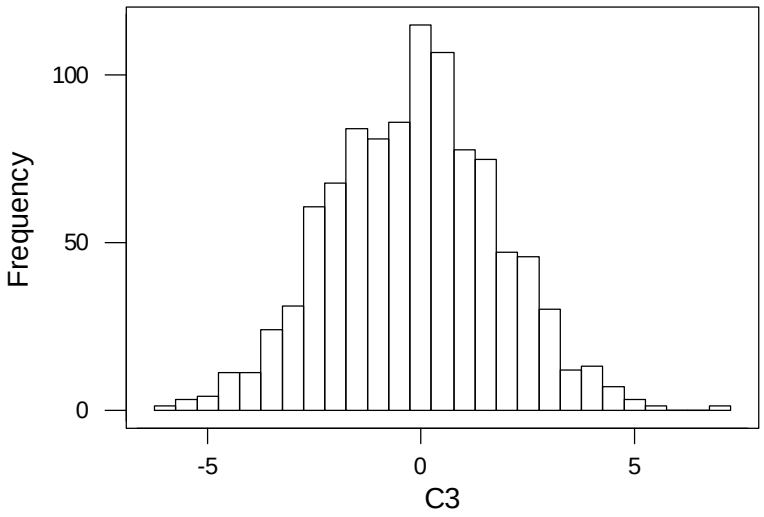
In contrast to the situation with thallium, there is very strong evidence in the data for technetium that the population means differ.

Question 2

The data generation follows the same pattern as above, with the entry in the Session Window being

```
MTB > %sampmns 50 50 1100 1000 c1
Executing from file: D:\MTBWIN\MACROS\sampmns.MAC
MTB > endmacro
MTB > %sampmns 50 50 1400 1000 c2
Executing from file: D:\MTBWIN\MACROS\sampmns.MAC
MTB > endmacro
MTB > Let c3 = c1-c2
```

The histogram of C3, the column containing the differences in sample means is shown below.

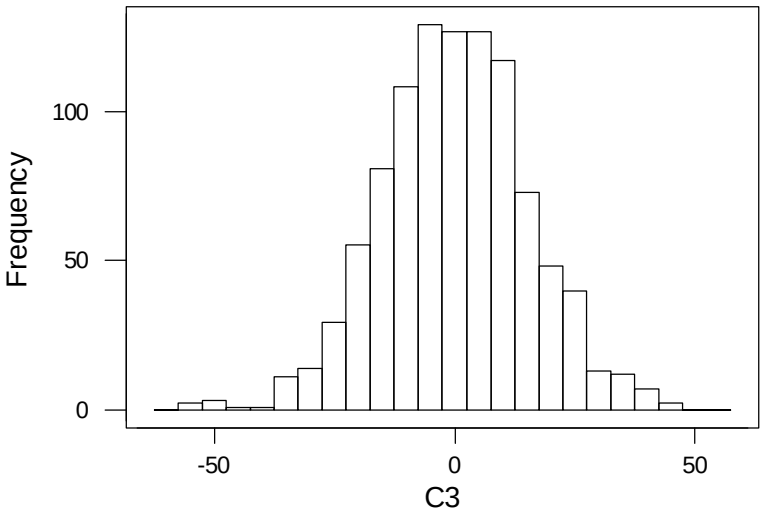


This is much less dispersed around 0 than the histogram for samples of size 11 and 14 which constituted Figure 1 in the study document. If the observed difference of 8.3 had arisen as the difference between samples of size 1100 and 1400, then this histogram shows that it would have been a very unusual occurrence *if the population means had been the same*. Counting the observations reveals that all 1000 differences in means are between –8.3 and 8.3, so as in question 1, $P < 0.001$.

Thus the implications of a given observed difference depend heavily on the sizes of samples from which they are obtained. With small samples, such as 11 and 14, the observed difference in means will be very variable and could quite plausibly have come from populations which have the same true mean. However, for larger samples, such as 1100 and 1400, the observed value will be a very much better estimate of the true difference in means.

Question 3

Repeating the exercise in question 2, but generating C1 to be 1000 means of samples of size 11, and C2 to be 1000 means of samples of size 1400, gives the following histogram for the column of differences.



This histogram is clearly much more dispersed than that for samples of 1100 and 1400. It is clear that a difference of 8.3 observed on the basis of samples of 11 and 1400 is certainly compatible with the population means being the same. This is confirmed by counting the number of observations in the relevant intervals. Using the **Tally...** command shows that of the 1000 means 288 are below –8.3, 287 are above 8.3, giving P = (288+287)/1000 = 0.575.

This example embodies an important lesson. The precision of the difference in two means is determined by the precision of both of them. You cannot make a difference arbitrarily precise by increasing the total number of observations if the number of observations made on one of the means remains small. For example, if you wish to compare some aspect of patients with a rare tumour with controls who do not have the tumour, then you cannot compensate for small numbers of cases by vastly increasing the number of controls.

End of solution sheet