Research Methods 2

Week 5: Exercise Sheet 1

Solution sheet

Question 1

The first ten numbers generated are

6.52505	5.39761
3.03954	5.99458
6.74382	6.68125
4.92357	4.59189
6.17919	4.56299

So 5 of these numbers lie between 4 and 6 (*remember that the ten numbers you generate will not necessarily be the same as these*).

For a Normal distribution with population mean 5 and population SD 1, the 68-95-99.7 rule implies that 68% of its values lie between 4 and 6 (i.e. within 1 SD of the mean). Now a representative sample from this population should reflect this property. So, in a sample of size 10 about 68% of the numbers will be within these limits, i.e. typically there will be about 7 of the values within these limits. However, the play of chance means that this value will not be obtained in every sample, sometimes it will be larger, sometimes, as here, smaller. However, it is plausible that in the long run, the number of values between 4 and 6 should, in some sense, 'average out' at just under 7.

If you repeat the exercise you obtain,

4.93363	4.56468
5.88620	5.89935
5.66039	2.68548
5.05077	5.51830
2.82105	4.01443

This time there are 8 values between 4 and 6.

The next two samples give

5.48602	5.54851
4.77882	5.60933
5.01522	5.07063
4.97880	6.25744
5.39768	5.60472

5.11788	6.59986
5.08650	4.96698
2.84947	5.62590
5.08801	4.60350
4.06913	6.65738

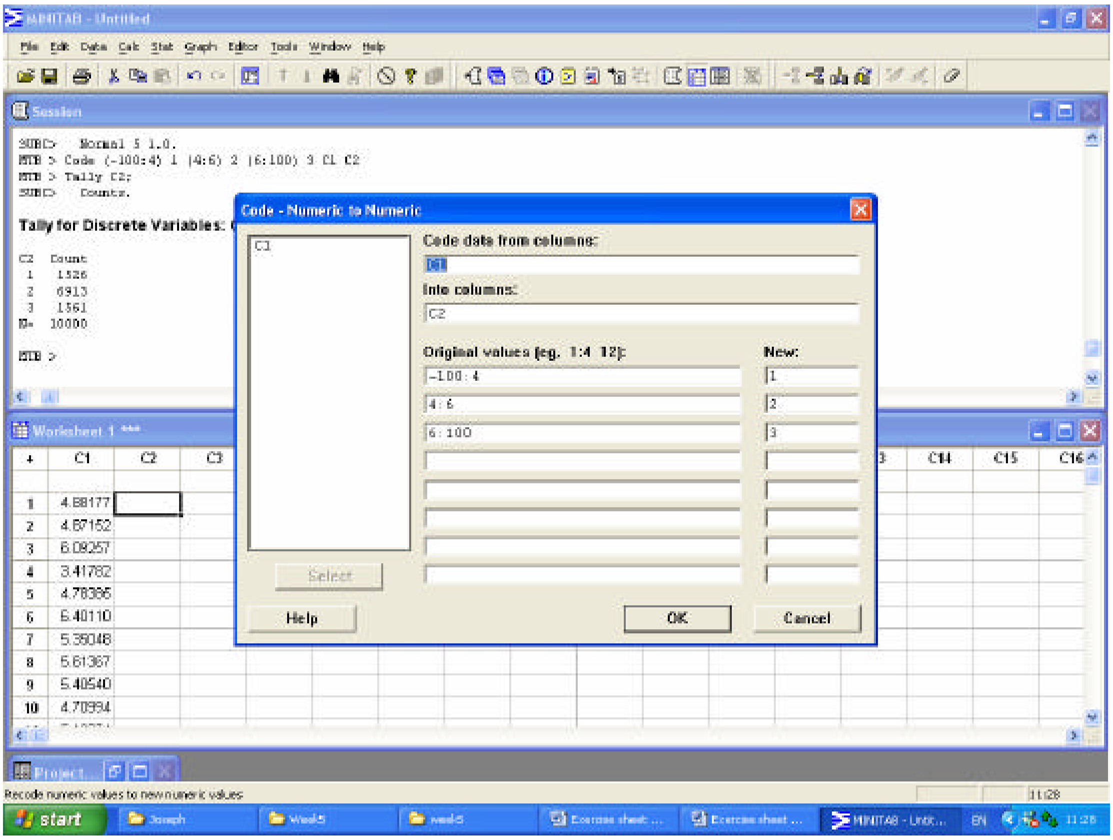
Which have, respectively, 9 and 7 values between 4 and 6.

If you pool these values, to effectively obtain a sample of size 40, the number of values between 4 and 6 is 5 + 8 + 9 + 7 =29. Now 29 out of 40 is 72.5%, which is not far from the 'expected' proportion of 68%.

Although the numbers you generate will be different, the same broad picture should emerge. The number of numbers between 4 and 6 varying substantially from 7 in the samples of size 10, but the proportion of the combined sample lying in this interval is closer to 0.68.

Question 2

Once the sample of size 10000 has been generated and placed in C1 (you could have used any spare column but in the screens shown below C1 has been used), you need to count how many lie between 4 and 6. This is done using the second method outlined in the hint to this question. This method first generates a second column (here C2 is used) which comprises just the values 1, 2 or 3. A value 1 is given if the value in the corresponding row of C1 is below 4, a value 2 if it is between 4 and 6 and a value 3 if it exceeds 6. This can be done by choosing **Code** *from the* **Manip** *menu and selecting the* **Numeric to Numeric.**. *option*. When completed as described in the hint the dialogue box looks as follows.



```
SUBC>   Normal 5 1.0.
MTB > Code (-100:4) 1 (4:6) 2 (6:100) 3 C1 C2
MTB > Tally C2;
SUBC>   Counts.

Tally for Discrete Variables: C

C2  Count
 1   1526
 2   6913
 3   1561
N=  10000

MTB >
```

Code - Numeric to Numeric

C1

Code data from columns: C1

Into columns: C2

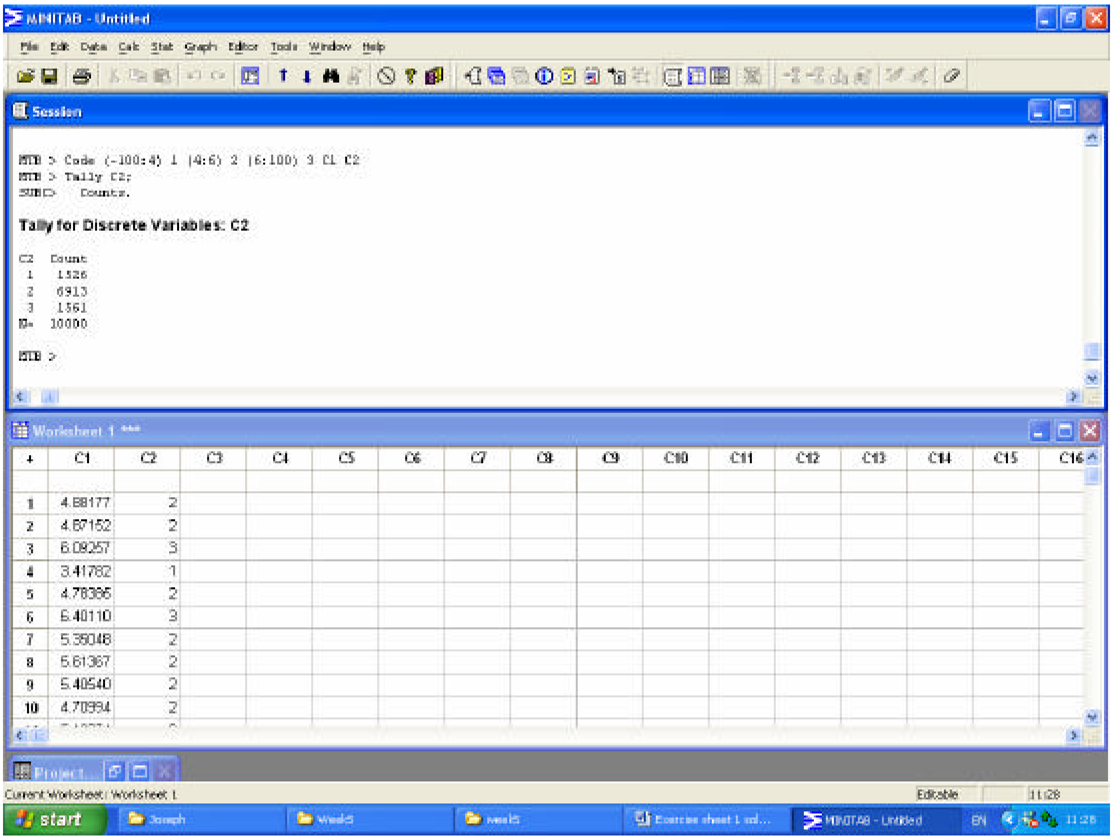
Original values (eg. 1:4 12):	New:
-100 : 4	1
4 : 6	2
6 : 100	3

Select   Help   OK   Cancel

	C1	C2	C3
1	4.88177		
2	4.87152		
3	6.08257		
4	3.41782		
5	4.78386		
6	6.40110		
7	5.35048		
8	5.61367		
9	5.40540		
10	4.70994		

Recode numeric values to new numeric values

Once this column has been constructed the number of 1s, 2s and 3s can be found using Tally command which is obtained from the **Stat** menu and then **Tables** and then **Tally Individual Variables...**. This gives the following screen.



```
MTB > Code (-100:4) 1 (4:6) 2 (6:100) 3 C1 C2
MTB > Tally C2;
SUBC>   Counts.
```

Tally for Discrete Variables: C2

```
C2  Count
 1   1526
 2   6913
 3   1561
N=  10000

MTB >
```

↓	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16
1	4.88177	2														
2	4.67152	2														
3	6.09257	3														
4	3.41782	1														
5	4.78386	2														
6	6.40110	3														
7	5.35048	2														
8	5.61367	2														
9	5.40540	2														
10	4.70994	2														

Current Worksheet: Worksheet 1

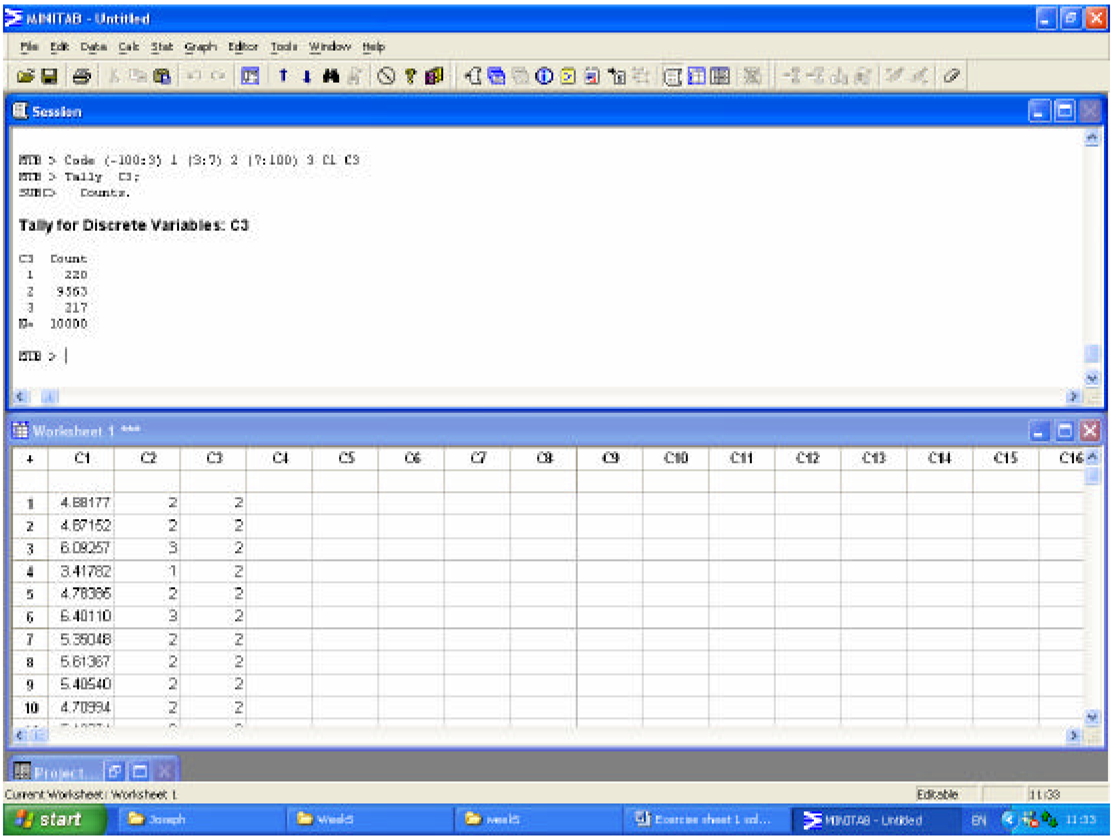
Editable

It can be seen from here that 6913 numbers in the sample are between 4 and 6, 1526 are less than 4 and 1561 are above 6.

Thus the percentage of the sample with 1 SD of the mean, 69%, is close to the expected value of 68%. It should also be noted that of the 3087 outside this range, approximately equal numbers lie above 6 and below 4, as would be expected from the symmetry of the Normal distribution.

Notice also how the proportion within 1 SD of the mean is much closer to its expected value when a very large sample is used. The importance of this feature will be explained in week 6.

Repeating the above coding exercise but with ‘3’ replacing ‘4’ and ‘7’ replacing ‘6’ and placing the results in column C3 allows us to find out how many of the sample lie within 2 SDs of the mean. The results of the tally command shown below reveals that 9563, i.e. approximately 95% of the sample lie within 2 SDs of the mean, confirming the ‘95’ part of the 68-95-99.7 rule. Also, almost equal proportions (i.e. 2.5%) lie above and below these limits.

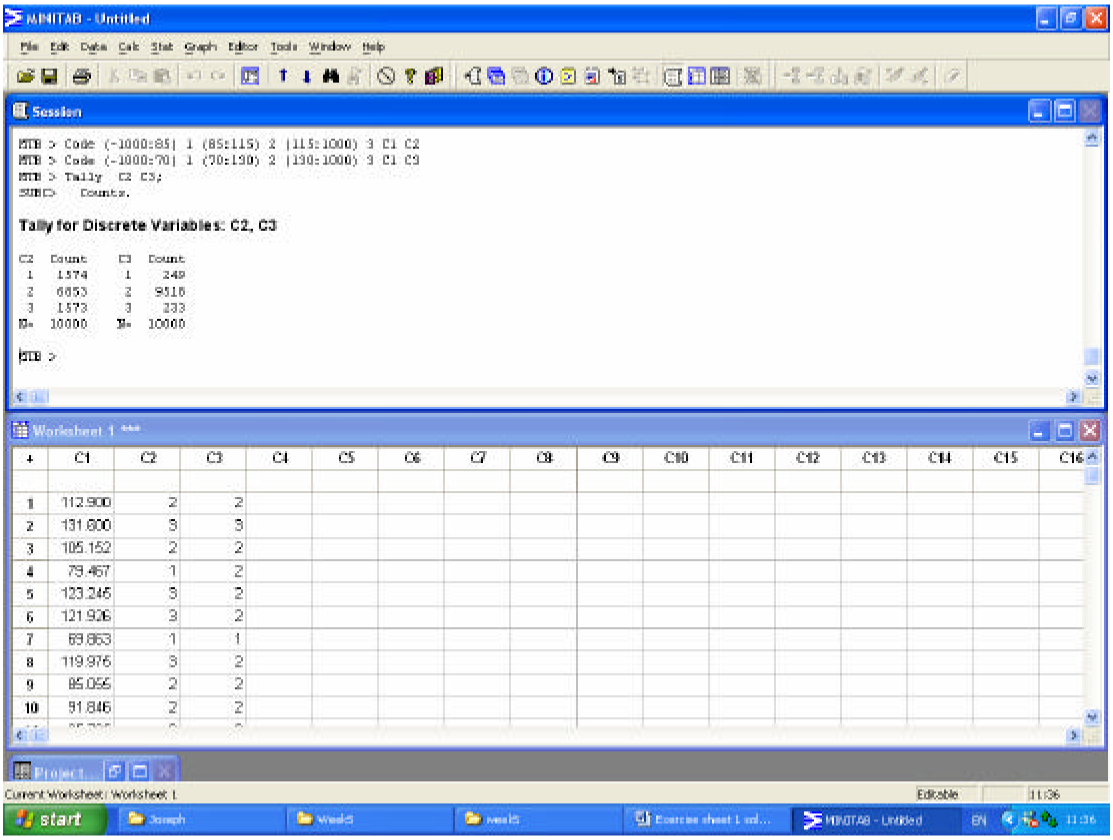


```
MTB > Code (-100:3) 1 (3:7) 2 (7:100) 3 C1 C3
MTB > Tally  C3;
SUBC>   Counts.

Tally for Discrete Variables: C3

C3  Count
 1    220
 2   9563
 3    217
N=  10000

MTB > |
```

+	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16
1	4.88177	2	2													
2	4.87152	2	2													
3	6.08257	3	2													
4	3.41782	1	2													
5	4.78386	2	2													
6	6.40110	3	2													
7	5.35048	2	2													
8	5.61367	2	2													
9	5.40540	2	2													
10	4.70994	2	2													

Question 3

Repeating question 2, but with a mean of 100, an SD of 15 and limits of 85 and 115 and of 70 and 130 gives the following results. Column C2 gives the counts in the intervals up to 85, 85 to 115 and over 115, whereas column C3 gives the counts in the intervals up to 70, 70 to 130 and over 130.



```
MTB > Code (-1000:85| 1 (85:115) 2 |115:1000) 3 C1 C2
MTB > Code (-1000:70| 1 (70:130) 2 |130:1000) 3 C1 C3
MTB > Tally  C2 C3;
SUBC>   Counts.
```

Tally for Discrete Variables: C2, C3

```
C2  Count    C3  Count
 1   1574     1    249
 2   6853     2   9518
 3   1573     3    233
N=  10000    N=  10000

MTB >
```

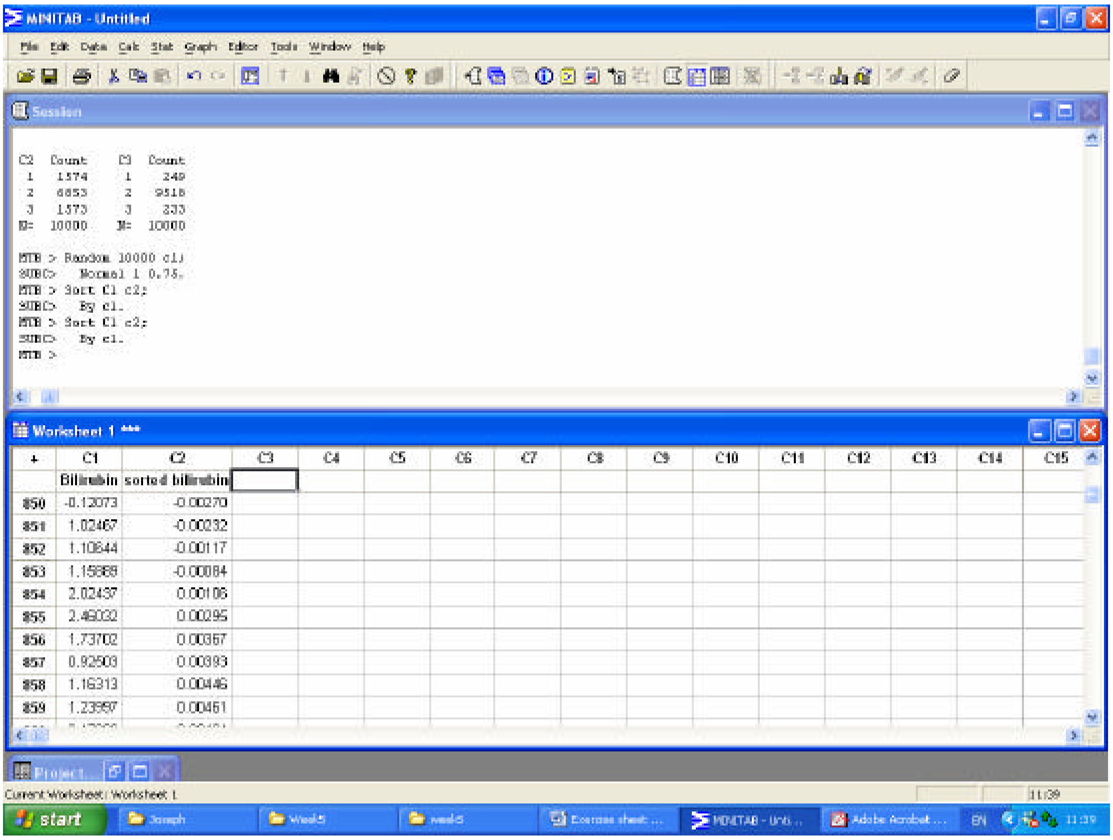
+	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16
1	112.900	2	2													
2	131.600	3	3													
3	105.152	2	2													
4	79.467	1	2													
5	123.246	3	2													
6	121.926	3	2													
7	69.863	1	1													
8	119.976	3	2													
9	85.056	2	2													
10	91.846	2	2													

The results are very similar to those obtained in question 2, with approximately 68% within 1 SD and 95% within 2 SDs of the mean, and with the remaining values nearly equally split above and below these intervals. This illustrates the fact that the 68-95-99.7 rule does indeed apply regardless of the particular mean and SD used.

Question 4

Using the technique used in all the previous questions it is straightforward to generate the 10000 bilirubin measurements. In order to count how many are negative it is possible to adapt the previous coding technique, by preparing a coded column with just two values. The first value corresponds to values in the interval –100:0 and the second to those in the interval 0:100. Alternatively the column of artificial bilirubin values could be sorted using the method outlined in a hint to the first Exercise Sheet in week 3.

Sorting the column and scrolling down the Data Window until you see both negative and positive values in the sorted column gives the following:



```
C2   Count     C3   Count
 1    1374      1     249
 2    6853      2    9518
 3    1573      3     233
N=   10000     N=   10000

MTB > Random 10000 c1;
SUBC>   Normal 1 0.75.
MTB > Sort C1 c2;
SUBC>   By c1.
MTB > Sort C1 c2;
SUBC>   By c1.
MTB >
```

↓	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15
	Bilirubin	sorted bilirubin													
850	-0.12073	-0.00270													
851	1.02467	-0.00232													
852	1.10644	-0.00117													
853	1.15869	-0.00084													
854	2.02437	0.00106													
855	2.49032	0.00295													
856	1.73702	0.00367													
857	0.92503	0.00393													
858	1.16313	0.00446													
859	1.23957	0.00461													

From this screen you can see that the numbers in C2 change sign from row 853 to 854. Therefore, as all the values in rows 1-852 are smaller than the value in row 853 (the column is now in ascending order), it follows that 853 of the 10000 values are negative.

This is perfectly reasonable for an abstract Normal population but it poses problems if you wish to use this Normal distribution to describe a variable, bilirubin concentration, which *cannot* be negative. What you have shown in this question is that the assumption that bilirubin concentration has a Normal distribution with mean 1 mg/l and SD 0.75 mg/l carries with it the implication that about 9% of the population will have negative values. As this is questionable, this assumption is untenable.

This will happen with any Normal distribution in which the SD is close in size to the mean. For example, the 68-95-99.7 rule implies that 16% of the population lies below μ-σ (make sure you understand this), and if, e.g., the mean is equal to the SD, then this is the same as saying that 16% of the population have negative values. For variables that are necessarily positive (and many encountered in medicine are) this implies that a Normal distribution is unlikely to hold.

What is the practical value in this observation? It is that for positive variables you can see from looking at the relative sizes of the mean and SD whether or not a Normal distribution is plausible. If the mean is larger than twice the SD then a Normal distribution *might* be OK. If the mean is less than the SD then a Normal distribution is unlikely to obtain. For means between one and two SDs, then the judgment is a finer one and a Normal distribution may or may not be acceptable.

The full assessment of whether or not a variable has a Normal distribution is an important question but one which is beyond the scope of this course. An obvious way to assess Normality is to plot a histogram but you need the full dataset to do this. If you are reading a paper and all that you have is the sample mean and SD of a positive variable, then comparing their sizes can give useful clues.

End of solution sheet