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.

4.93363	4.56468
5.88620	5.89935
5.66039	2.68548
5.05077	5.51830
2.82105	4.01443

If you repeat the exercise you obtain,

This time there are 8 values between 4 and 6.

The next two samples give

.48602	5.54851	5.117	6.5998
4.77882	5.60933	5.086	4.96698
5.01522	5.07063	2.849	947 5.62590
4.97880	6.25744	5.088	4.60350
5.39768	5.60472	4.069	913 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.

≥n	IITAB - Un	tilled									- @ X
Ma	Edit Data	Cek Sta	e graph t <u>d</u>	tor Italii Window Hei	Þ						
			nal	TIAR)	0 ? 💷 🛛 🖸	🖄 🛈 🖸 🗟 🛍 🖏 🗍	C 📺 🌃 🕅 -1 -	5 da 68 🖉	1.0		
Q 54											
1977)9 1977)9 1977)9 1977)9	> Tally i	100:4) 2; z.	1 (4:6) 2	(6:100) 3 CL C2 Code - Numeric to No		om columns:		×			5
1	L526			100	161						
2	6913				Inte celumn	st					
	1361				CS						
ITTE					Original val	ues (eg. 1:4-12):	New:				9
	50.				-100:4		1	_			~
4	AL I				4.6		2	_			2
₩	oriksheet 1		_		6:100		3				🛛
4	C1	C2	C3		-			3	C14	C15	C16 🗠
1	4.68177	19 - 14 19 - 14								-	
2	4.67152	-						-			
3	6.09257			<u> </u>			-	1			
4	3.41782			Select			-				
5	4.78395				4			1			
6	6.40110			Help		0	K Can	cel			
1	5.39048		-					102	-	-	
8	5.61367								-		
10	4,70994							-	-	-	
eiii	- 10071		1						1	-	
-	0.025 19		3								
	opert. E		uneric values								1/28
-	tart			b West	Converto	Exercise sheet	Ecores sheet	та милтае	-LINE	EN CIT	
					and the second	and the second s	and a construction of the second			Contraction of the	

Once this column has been constructed the number of 1s, 2s and 3s can be found using Tally command which is obtained from the **<u>S</u>tat** menu and then **<u>T</u>ables** and then **<u>T</u>ally Individual Variables...** This gives the following screen.

1 5 578 578	y for Disc	100:4) 1 2: z.	(4:6) 2	(6:10D) 3		0 ? Ø	1	i () I	i 1 1 3			-1-20	18 12	d. 0	6	. 6
178 178 208 7all	> Code (- > Tally C D Count y for Disc Count	2; z.			1 CL C2											
718 718 72 1	> Tally C D Count y for Disc Count	2; z.			i CL C2											
2	Count	rete Varia	bles: C	2												
L.																
3	1526 6913 1551 10000															
	201252															
B	8) (A															
	A.															2
v	orisheet 1	***													8	. 0
ŧ.	C1	C2	C3	C4	CS	06	a	08	CB	C10	C11	C12	.013	C14	C15	CI
	100000000	ci-														
1	4.E8177	2														
ł.	4.87152	2														
	6.09257	3														
1	- 17 - 21 - 21 - 21 - 21 - 21 - 21 - 21															
-	3.41782	1														
5	3.41782 4.78386	1														
5	3.41782 4.78385 6.40110	1 2 3														
4 5 6 7	3.41782 4.78386 6.40110 5.35048	1 2 3 2														
4 5 6 7 8	3.41782 4.78385 6.40110 5.39048 5.61367	1 2 3 2 2														
5 5 7 3	3.41782 4.78385 5.40110 5.39048 5.61367 5.40540	1 2 3 2 2 2														
5 5 7 8 9	3.41782 4.78365 6.40110 5.39048 5.61367 5.40540 4.70994	1 2 3 2 2 2 2 2 2														
5 7 8 0	3.41782 4.78385 5.40110 5.39048 5.61367 5.40540	1 2 3 2 2 2														6
3 4 5 6 7 8 9 10	3.41782 4.78365 6.40110 5.39048 5.61367 5.40540 4.70994	1 2 3 2 2 2 2 2														

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.

_	Edit Data															
3		22 🚳	06	114	MA	0 ? 🗗	10	i 🛈 🖸	10		圖風	-2-20	日间间	1.0		
50	ssion															
au a	> Code (- > Tally D Dount	C3;	3:7) 2	7:100) :	3 CL C3											
rall	y for Disc	ete Varia	bles: C3													
1 2 3 9+	Count 220 9563 217 10000															
TE	21															
	8.0															
	A															
-				_												_
-	orksheet 1	***													. 8	. 0
-	orisheet 1 C1	••• C2	cı	сі	cs	06	a	08	cy	.C10	C11	C12	C13	CH	C15	
] w ↓	CI	C2	-	сі	cs	06	a	08	C	C10	CII	C12	C13	C14		
• •	C1 4.68177	2	2	сі	cs	06	a	08	C	C10	C11	C12	C13	C14		
1 2	C1 4.68177 4.67152	C2 2 2	2	с	cs	C6	a	08	C	C10	CII	C12	C13	C14		
1 2 3	C1 4.88177 4.67152 6.09257	C2 2 2 3	2 2 2	сı	cs	C6	a	08	C3	C10	CII	C12	C13	C14		
1 2	C1 4.68177 4.67152	C2 2 2 3 1	2 2 2 2 2	cı	cs	06	a	08	C	C10	C11	C12	C13	C14		
1 2 3 4	C1 4.88177 4.87152 6.08257 3.41782	C2 2 2 3	2 2 2 2 2 2 2	с	cs	C6	a	08	C	C10	C11	C12	C13	C14		
+ 1 2 3 4 5	C1 4.88177 4.67152 6.09257 3.41782 4.78366	C2 2 2 3 1 2	2 2 2 2 2	а	cs	C6	a	03	G	C10	C11	C12	C13	C14		
+ 1 2 3 4 5 6	C1 4.E8177 4.67152 6.08257 3.41782 4.78366 6.40110	C2 2 3 1 2 3 3 3 3	2 2 2 2 2 2 2 2 2	сі	CS	06	a	03	G	.C10	CH	C12	013	C14		
4 1 2 3 4 5 6 7	Ct 4.88177 4.67152 6.09257 3.41782 4.78385 6.40110 5.35048	C2 2 3 1 2 3 2 3 2 2	2 2 2 2 2 2 2 2 2 2 2 2 2	Cł	C	C6	a	08	0	C10	CH	C12	613	СН		Cit
4 1 2 3 4 5 6 7 8	C1 4.68177 4.67152 6.09257 3.41762 4.78366 6.40110 5.39048 5.61367	C2 2 3 1 2 3 2 2 2 2	2 2 2 2 2 2 2 2 2 2 2 2 2 2	Cł	G	C6	a	08	0	C10	CH	C12	C13	СН		
+ 1 2 3 4 5 6 7 8 9 10	C1 4.88177 4.67152 6.09257 3.41782 4.78366 6.40110 5.39048 5.61367 5.40540	C2 2 3 1 2 3 2 2 2 2 2	2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2	а	G	C6	a	3	G	.C10	Cfi	C12	C13	СН		
+ 1 2 3 4 5 6 7 8 9 10	C1 4.88177 4.87152 6.09257 3.41782 4.78365 6.40110 5.38048 5.61367 5.40540 4.70394 4.70394	C2 2 2 3 1 1 2 3 2 2 2 2 2 2 2 2 2 2 2 2	2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2	а	CS	C6	9	3	g	C10	Cfi	C12	C13	C14		Cl
+ 1 2 3 4 5 6 7 8 9 10	C1 4.88177 4.67152 6.09257 3.41782 4.78366 6.40110 5.39048 5.61367 5.40540 4.70994	C2 2 3 1 2 3 2 2 2 2 2 2 2 2 2 2 2 2 2 2	2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2	64	CS	C6	9	03	C	C10	CH	C12	C13	C14	CIS	C

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.

	45				Window Hel	P										
		四國	e) es 🚺	1 1 4	AR	0 ? 🗗	100	0 🖸 🖸	a		圖圖	-2-20	自同型	1. 1		
50	ssion															
ET.	> Code (- > Tally															
all	y for Disc	rete Varia	bles: C2,	C3												
1 2 3	1374 6853 1373	C3 Count 1 249 2 9516 3 233 3-10000														
TB	2															
34																2
-																
-	orksheet 1															. 0
w		•••	cı	C4	cs	C6	a	CB4	cy	C10	C11	C12	613	CI4	C15	
w	orisheet 1 C1	C2		Cł	cs	C6	a	08	cı	C10	CII	C12	C13	CI4		
w	orksheet 1	C2 2	2	Cł	cs	C6.	a	08	cy	C10	CII	C12	C13	C14		
W	orischeret 1 C1 112.900	C2		с	cs	C6	a	03	0	C10	C11	C12	C13	C14		
W	C1 C1 112,900 131,000	C2 2 3	2 3 2	сі	cs	C6	a	08	Q	C10	CII	C12	C13	C14		
W	112.900 1131.600 105.152	C2 2 3 2	2 3 2 2	Cł	cs	C6	a	03	0	C10	C11	C12	C13	C14		
W	CI 112.900 131.000 105.152 79.467	C2 2 3 2 1	2 3 2	с	cs	C6	a	03	G	C10	CII	C12	C13	C14		
W 1 2 3 4 5 5	112.900 131.900 105.152 79.467 123.245	C2 2 3 2 1 3	2 3 2 2 2	с	cs	C6	a	08	G	C10	CII	C12	C13	C14		
1 2 3 4 5 5 7	Ct Ct 112.900 131.600 105.152 73.467 123.245 121.926	C2 2 3 2 1 3 3 3 1 3	2 3 2 2 2 2 2 1 2	сі	cs	06	<i>a</i>	03	0	.C10	CH	C12	C13	CH		
-	112.900 131.600 105.152 73.467 123.245 121.926 69.853	C2 2 3 2 1 3 3 3 1 3 2	2 3 2 2 2 2 1 1 2 2	Ci	C	C6	a	08	()	CIO	CH	C12	613	СН		C1
1 2 3 4 5 5 7 8	112.900 131.600 105.152 79.467 123.245 121.926 69.853 119.976	C2 2 3 2 1 3 3 3 1 3	2 3 2 2 2 2 1 2 2 1 2 2 2 2 2 2 2 2	Cł	G	C6	a	08	0	C10	CH	.012	C13	CH		
	112 900 131 800 105 152 73 467 123 246 121 936 69 863 119 976 85 056	C2 2 3 2 1 3 3 3 1 3 2	2 3 2 2 2 2 1 1 2 2	с	G	C6	a	3	0	C10	CH	.012	C13	CH		CI
W F F F F F F F F F F F F F F F F F F F	112,900 131,000 105,152 73,467 123,246 121,926 63,653 119,976 85,055 91,846	C2 3 2 1 3 3 1 3 3 2 2 2 2 2 2 2 2 2 2 2	2 3 2 2 2 2 1 2 2 1 2 2 2 2 2 2 2	с	C	06	9	3	g	.C10	CH	C12	C13	CH		

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:

1 3 9= 1 900 > 900 > 900 > 900 > 900 > 900 >	ount C 1574 : 6853 : 1573 : 0000 M Boneal : Sort Cl : Sort Cl : Sort Cl :	2 9518 3 233 = 10000 0000 cl) 1 0.75, 22;													
1 2 3 3 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9 9	1574 6653 1573 0000 M Random D Wormal Sort C1 Sort C1 Sort C1	1 249 2 9518 3 233 = 10000 0000 cLJ 1 0.75. c2;													
 30100 	Normal Sort Cl (By Cl. Sort Cl (1 0.75, :2;													
<pre>state</pre>															
-	rksheet 1 *	6													
+	C1	2	C3	C4	C5	CG	C7	C8	C9	C10	C11	C12	C13	C14	C15
	Bilirubin s	sorted bilirubin													
150	-0.12073	-0.00270		1											
51	1.02467	-0.00232													
	1.100.11	-0.00117													
52	1.10644														
-	1.10644	-0.00084													
53															
153 154	1.15889	-0.00084													
153 154 155	1.15889 2.02437	-0.00084 0.00106													
152 153 154 155 156 157	1.15889 2.02437 2.46032	-0.00084 0.00106 0.00295													
153 154 155 156	1.15889 2.02437 2.46032 1.73702	-0.00084 0.00106 0.00295 0.00367													
153 154 155 156 157	1.15889 2.02437 2.46032 1.73702 0.92503	-0.00084 0.00106 0.00295 0.00367 0.00393													

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