Guatemalan cholesterol example

(last update 9 January 2018)

This example is taken from Devore and Peck, *Statistics*, 3 ed., (1997), Duxbury, p. 23. The original source is "The Blood Viscosity of Various Socioeconomic Groups in Guatemala" in *The American Journal of Clinical Nutrition*, Nov., 1964, 303–307. The Institute of Nutrition of Central America and Panama measured the serum total cholesterol levels for a group of 49 adult, low–income rural Guatemalans and for a group of 45 adult, high–income urban Guatemalans. These serum total cholesterol levels (in mg/dL) are provided in Table 1.

Table 1. Guatemalan cholesterol data.

Rural group cholesterol levels (in mg/dL).

95	108	108	114	115	124	129	129	131	131
135	136	136	139	140	142	142	143	143	144
144	145	146	148	152	152	155	157	158	158
162	165	166	171	172	173	174	175	180	181
189	192	194	197	204	220	223	226	231	

Urban group cholesterol levels (in mg/dL).

133	134	155	170	175	179	181	184	188	189
190	196	197	199	200	200	201	201	204	205
205	205	206	214	217	222	222	227	227	228
234	234	236	239	241	242	244	249	252	273
279	284	284	284	330					

Before we compute any summary statistics consider the histograms provided below. There are three pairs of histograms. The first two figures contain proper relative frequency histograms while the third contains stem and leaf histograms. Based on these histograms we can see that both of these cholesterol level distributions are basically mound shaped with some skewness to the right. In the rural group there are four individuals with somewhat high cholesterol levels (220 or more); there is a gap of 16 separating the cholesterol levels of these individuals from the rest of the rural group. It is this group of four observations which causes the rural distribution to appear skewed to the right. The urban group has similar slightly unusual groups of cholesterol levels; one group having somewhat low levels and one having somewhat high levels. There is one unusually large value (330) in the urban group that we might consider an outlier, since there is a gap of 46 between 330 and the next largest value. (An outlier is an observation that is widely separated from the majority of a distribution.) We will need to consider the implications of this outlier in our analysis of this example. Note that without the urban outlier (330), the cholesterol distribution of the urban group is essentially symmetric.

It is also apparent that the people in the urban group tend to have higher cholesterol levels than the people in the rural group.

There appears to be more variability among the cholesterol levels of the people in the urban group. With the urban outlier (330) there seems to be much more variability in the cholesterol levels of the people in the urban group. Without this outlier, there appears to be only slightly more variability in cholesterol levels of the people in the urban group.

Figure 1. Guatemalan cholesterol level histograms.

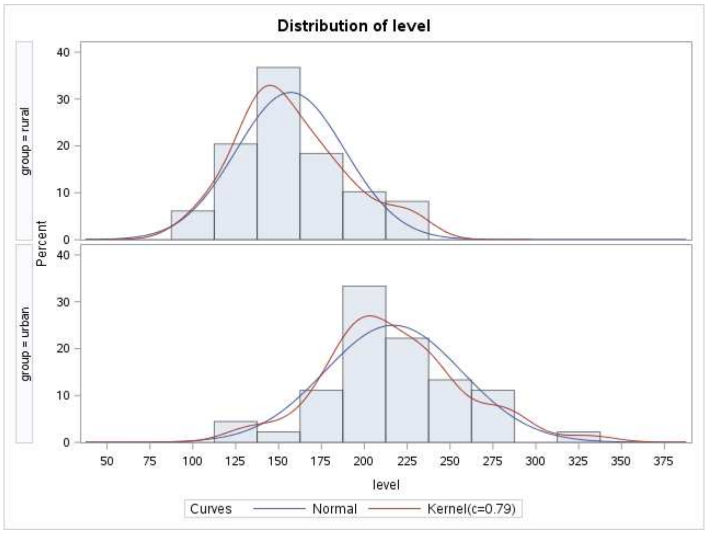


Figure 2. Guatemalan cholesterol level histograms.

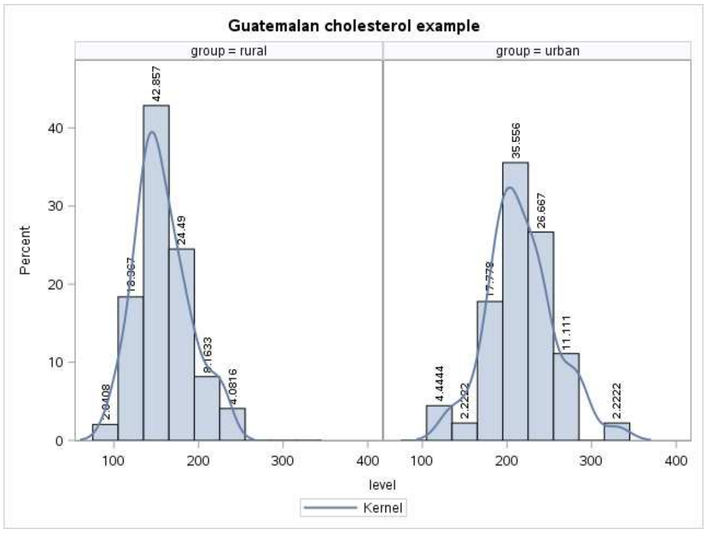


Figure 3. Guatemalan cholesterol level stem and leaf histograms.

The stem represents tens and the leaf represents ones. (mg/dL)

Rural		Urban	
9	5	9	
10	88	10	
11	45	11	
12	499	12	
13	115669	13	34
14	0223344568	14	
15	225788	15	5
16	256	16	
17	12345	17	059
18	019	18	1489
19	247	19	0679
20	4	20	001145556
21		21	47
22	036	22	22778
23	1	23	4469
24		24	1249
25		25	2
26		26	
27		27	39
28		28	444
29		29	
30		30	
31		31	
32		32	
33		33	0

The five number summaries and the associated distances based on them are provided, for the rural group, for the entire urban group, and for the urban group omitting 330, in Table 2. The steps involving in computing the medians and quartiles, for the rural group and the entire urban group, are outlined below.

For the rural group there are $n = 49$ observations so that

(1) $49/2 = 24.5$, thus the median 152 is obs. no. 25, corresponding to the first 2 leaf in the 15 stem.

(2) $49/4 = 12.25$, thus the first and third quartiles are $Q_1 = 136$, the $13th$ observation counting up, corresponding to the second 6 leaf in the 13 stem, and $Q_3 = 174$, the $13th$ observation counting down, corresponding to the second 4 leaf in the 17 stem.

For the urban group there are $n = 45$ observations so that

(1) $45/2 = 22.5$, thus the median 206 is obs. no. 23, corresponding to the 6 leaf in the 20 stem.

(2) $45/4 = 11.25$, thus the first and third quartiles are $Q_1 = 196$, the $12th$ observation counting up, corresponding to the 6 leaf in the 19 stem, and $Q_3 = 239$, the $12th$ observation counting down, corresponding to the 9 leaf in the 23 stem.

Table 2. Five number summaries with distances.

Rural group. (mg/dL) n=49					
min:	95				
Q_1:	136	Q_1- min:	41	med - min:	57
med:	152	med - Q_1:	16		
Q_3:	174	Q_3- med:	22	max - med:	79
max:	231	max - Q_3:	57		
Urban group (all). (mg/dL) n=45					
min:	133				
Q_1:	196	Q_1- min:	63	med - min:	73
med:	206	med - Q_1:	10		
Q_3:	239	Q_3- med:	33	max - med:	124
max:	330	max - Q_3:	91		
Urban group (omit 330). (mg/dL) n=44					
min:	133				
Q_1:	193	Q_1- min:	60	med - min:	72.5
med:	205.5	med - Q_1:	12.5		
Q_3:	237.5	Q_3- med:	32	max - med:	78.5
max:	284	max - Q_3:	46.5		

Table 3. Summary statistics for the cholesterol example.

group	mean	median	std. dev.	range	IQR
rural	157.02	152	31.75	137	38
urban (all)	216.87	206	39.92	197	43
urban (omit 330)	214.30	205.5	36.42	151	42

Box plots for the Guatemalan cholesterol example are provided in Figures 4 and 5. These simple graphical displays give a visual impression of the information in Table 2 (the five number summary values and the distances among these values). A box plot does not convey as much information about the shape of a distribution as a histogram, but, it does give a useful graphical impression of the shape of the distribution (including skewness or symmetry). Box plots are particularly useful for quick comparisons of two or more distributions.

Figure 4. Guatemalan cholesterol level boxplots.

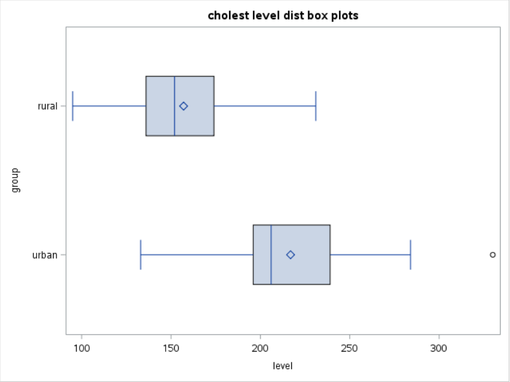
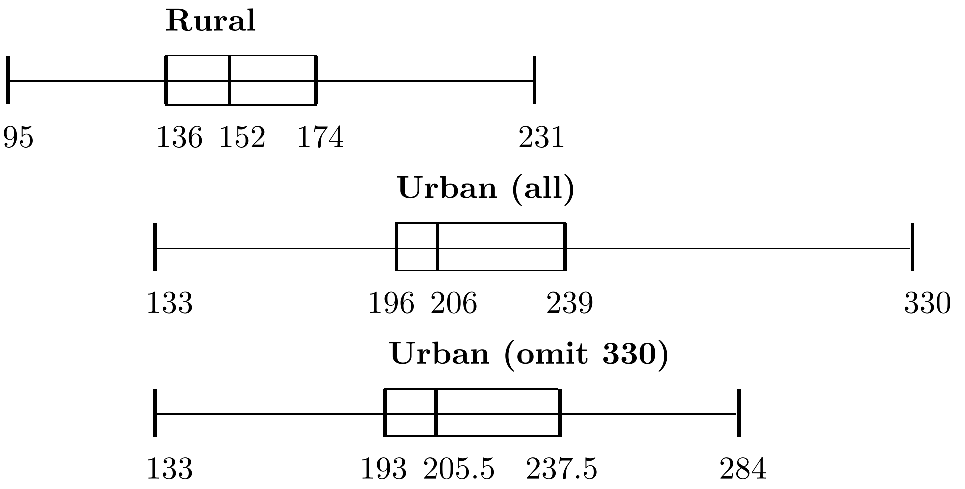


Figure 5. Box plots for cholesterol level.



Notice that each box plot has five vertical marks indicating the locations of the five number summary values. The box which extends from the first quartile to the third quartile and is divided into two parts by the median gives an impression of the distribution of the values in the middle half of the distribution. In particular, a glance at this box indicates whether the middle half of the distribution is skewed or symmetric and indicates the magnitude of the interquartile range (the length of the box). The line segments (whiskers) which extend from the ends of the box to the extreme values (the minimum and the maximum) give an impression of the distribution of the values in the tails of the distribution. The relative lengths of the whiskers indicate the contribution of the tails of the distribution to the symmetry or skewness of the distribution.

Returning to the cholesterol example first consider the shapes of the cholesterol distributions. We can use the distances, based on the five number summary, given in Table 2 (and shown in the box plots) to quantify the degree of skewness in these distributions.

First consider the cholesterol distribution for the rural group. The range of the right half of the distribution (max – med = 79) is greater than the range of the left half of the distribution (med – min = 57); Since 79 is about 40% larger than 57 (79/57 = 1.386), this comparison shows that, overall, the rural group cholesterol distribution is skewed to the right. Next consider the middle half (center) of the distribution, that is the part of the distribution between Q_1 and Q_3. The range of the right half of this central region (Q_3 – med = 22) is greater than the range of the left half of this central region (med – Q_1 = 16). Here, again, the range on the right 22 is about 40% more than the range on the left (22/16 = 1.375) indicating that the middle half of the rural group cholesterol distribution is skewed to the right. Finally, consider the tails of the distribution, that is the lower fourth, from the minimum to Q_1, and the upper fourth, from Q_3 to the maximum. The range of the right tail (Max – Q_3 = 57) is greater than the range of the left tail (Q_1 – min = 41). As before, the range on the right is about 40% more than the range on the left (57/41 = 1.390) indicating that there is skewness in the tails of the rural group cholesterol distribution. All of these comparisons support our contention that the cholesterol distribution for the rural group is skewed right and they also show that the degree of skewness in the tails is similar to the degree of skewness in the middle of the distribution.

Next consider the urban group, including the outlier. The range of the right half of the distribution (max – med = 124) is much greater than the range of the left half of the distribution (med – min = 73); Since 124 is about 70% larger than 73 (124/73 = 1.699), this comparison shows that, overall, the urban group cholesterol distribution is strongly skewed to the right. For the middle half (center) of the distribution, we find that the range of the right half of the central region (Q_3 – med = 33) is greater than the range of the left half of the central region (med – Q_1 = 10). Here the range on the right 33 is about 3 times the range on the left (33/10 = 3.3) indicating that the middle half of the urban group cholesterol distribution is very strongly skewed to the right. The skewness is less extreme in the tails of this distribution with the range of the right tail (Max – Q_3 = 91) being greater than the range of the left tail (Q_1 – min = 63). For the tails, the range on the right is about 40% more than the range on the left (91/63 = 1.444) indicating that there is skewness in the tails of the urban group cholesterol distribution but it is not as strong as the skewness in the center of the distribution. As with the urban group cholesterol distribution all of these comparisons support our contention that the cholesterol distribution for the urban group is skewed right. For the urban group we see that the degree of skewness in the tails is lower than the degree of skewness in the middle of the distribution.

If we omit the outlier (330) from the urban group, then the direction of skewness in the tails of the urban group cholesterol distribution reverses and overall the distribution appears reasonably symmetric. That is, without the outlier the range of the right half of the distribution (max – med = 78.5) is only slightly larger than the range of the left half of the distribution (med – min = 72.5); but, the range of the left tail (lower fourth) Q_1 – min = 60 is now larger than the range of the right tail (upper fourth) Max – Q_3 = 46.5.

With the outlier the range 197 for the urban group is much larger than the range 137 for the rural group. If we omit the outlier, then the range for the urban group is 151 which is still larger than 137 but not by so much. On the other hand, if we consider the interquartile ranges, 38 for the rural group and 43 (44.5 without the outlier) for the urban group, we find that there is a similar amount of variability in the middle halves of these distributions. Hence, our contention that there is much more variability among the cholesterol levels of the urban Guatemalans depends very heavily on the cholesterol level of one individual. Whether we include this individual or not, we are justified in claiming that there is more variability among the cholesterol levels of the urban Guatemalans.

Based on our analysis of these cholesterol level distributions we might propose several hypotheses or conjectures about why these distributions differ as they do. First we might conjecture that the rural Guatemalans are probably more physically active and eat food which is lower in fat than the urban Guatemalans. This would cause the rural Guatemalans to tend to have lower cholesterol levels. Second, we might argue that there is less variability in the cholesterol levels of the rural Guatemalans because their lifestyles and eating habits are probably quite similar.

In this example, if we base our comparisons of the location and the amount of variability in these distributions on the mean and standard deviation we reach essentially the same conclusions as we did when using the five number summary.