IT3030 Biostatistics Quiz#1 2019.02.26 ID:_____ Name:_____

1. (40%) Given the FEV1 data shown in Table 3.1. (a) Compute the mean value for the correct 13 observations (to 2 digits after decimal point, or 2D). (b) Compute the mean value for the 13 observations to 2D, using 40.2 as the 11th observation. (c) Rank these 13 values in both (a) and (b), from smallest to largest respectively, and determine the median for each of two sorted list. (d) State your observation/conclusion for whether the mean or median is better to predict the "center" for the sorted array. That is, which one is more resistant to a wrong entry such as 40.2 introduced as the 11th entry in terms of giving a better "center"?

Answer:

(a):

```
>> a=[2.3 2.15 3.5 2.6 2.75 2.82 4.05 2.25 2.68 3 4.02 2.85 3.38]; //the "correct" list, unsorted.
>> mean(a)
ans =       2.9500
```

(b):

```
>> b=a; b(11)=40.2; // The "incorrect" list, replacing the 11th entry to a new value.
>> mean(b)
ans =       5.7331
>>
```

(c):

```
>> sort(a)    // MATLAB function "sort"…
ans =
   Columns 1 through 8
      2.1500    2.2500    2.3000    2.6000    2.6800    2.7500    2.8200    2.8500
   Columns 9 through 13
      3.0000    3.3800    3.5000    4.0200    4.0500

>> median(a) // MATLAB function "median"…
ans =       2.8200

>> sort(b)
ans =
   Columns 1 through 8
      2.1500    2.2500    2.3000    2.6000    2.6800    2.7500    2.8200    2.8500
   Columns 9 through 13
      3.0000    3.3800    3.5000    4.0500   40.2000

>> median(b)
ans =       2.8200
>>
```

(d):

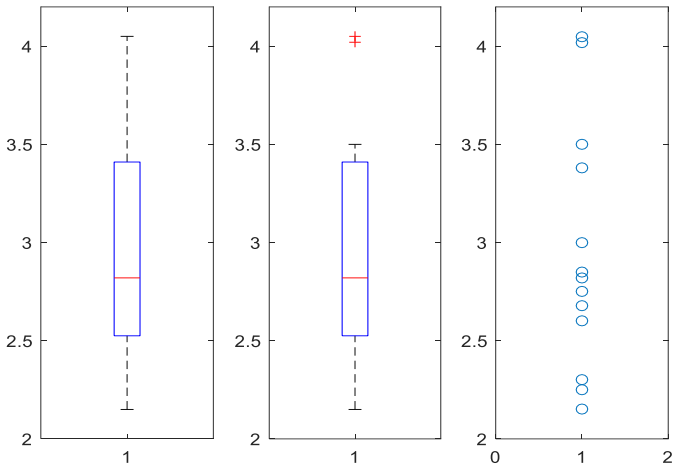
Both medians from the two arrays predicted 2.82, as seen from the two sorted arrays. The mean value from array "a" is 2.95, and from array "b" it is 5.7331. Apparently, the median predicted the "center" value better than the mean value did. In other words, mean value is more sensitive (or median is more resistant) should there be an error introduced into the list.

2. (40%) (a) Determine Q1, Q2 and Q3 for the 13 “correct” observations in Table 3.1. (b) Draw a box plot with its two whiskers. Use the box height for locating the two extreme values (the whiskers). (c) Do you have outliers (values that are more extreme than the whiskers)? If yes, what are they? (d) Redo (c) by using only half of the box height for locating the two whiskers. Any outliers?

Answer:
(a):
>> **quantile**(a, [.25 .5 .75]) // MATLAB function “quantile”…
ans = 2.5250 2.8200 3.4100

>> Upper=ans(3); Lower=ans(1);

>> IQR=**iqr**(a) // MATLAB function “iqr”…
IQR = 0.8850



(b):
>> Upper+IQR
ans = 4.2950

>> Lower-IQR
ans = 1.6400

Recall that the sorted array is [2.1500 2.2500 2.3000 2.6000 2.6800 2.7500 2.8200 2.8500 3.0000 3.3800 3.5000 4.0200 4.0500]. Therefore, the larger whisker would be the largest one no more than the upper bound 4.2950 obtained above, which is 4.0500. The smaller whisker would be the smallest one no less than the lower bound 1.6400 obtained above, which is 2.1500.

(c):
There is no outlier, since the whiskers represent the two ends of the array.

(d):
>> Upper+0.5*IQR
ans = 3.8525

>> Lower-0.5*IQR
ans = 2.0825

Therefore, the larger whisker would be the largest one no more than the upper bound 3.8525 obtained above, which is 3.5000. This gives both 4.0200 and 4.0500 outliers. On the other hand, the smaller whisker would be the smallest one no less than the lower bound 2.0825 obtained above, which is 2.1500. For this end, there is no outlier.

One can use MATLAB functions “subplot” and “boxplot” to do the boxplots for above two cases. In this case, three subplots were created for side-by-side comparison:
>> **subplot**(1,3,1);**boxplot**(a,'whisker',1.0);**ylim**([2 4.2])
>> **subplot**(1,3,2);**boxplot**(a,'whisker',0.5);**ylim**([2 4.2])
>> **subplot**(1,3,3);**plot**([1 1 1 1 1 1 1 1 1 1 1 1 1], **sort**(a), 'o');**ylim**([2 4.2])

Note: The whiskers extend to the most extreme data points not considered outliers, and the outliers are plotted individually using the '+' symbol.

3. (20%) (a) Create a frequency table for the 57 scores shown in class, using 0~20, 21~40, 41~60, 61~80 and 81~100 as intervals. (b) Draw the corresponding histogram, using relative frequency (rather than absolute frequency) for the y-axis labels.

Answer:
(a):

Category	Cnt
1: 0~20	1
2: 21~40	21
3: 41~60	21
4: 61~80	9
5: 81~100	5

(b):

```
>> F=[1 21 21 9 5];
>> FP=F/sum(F)
FP =
0.0175    0.3684    0.3684    0.1579    0.0877
>> bar(FP)
>>
```