

Suggested Answers for WarmUps for Lesson 04

Note: I do NOT expect students to have the same observations as mine, I just want them to make SOME observations.

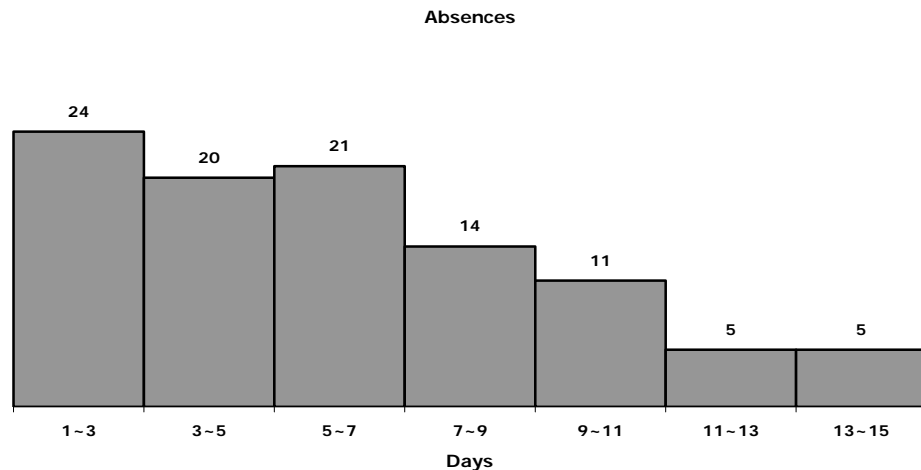
Absent

Mean number of days absent from work per year over a five year period for a sample of employees.

5.4	6.6	8.2	2.4	5.7	6.1	2.9	13.5	4.6	6.7
4.1	2.1	2.8	2.7	11.8	6	2.2	6.7	8.6	6.7
11.5	3.8	2.4	13.4	2.5	6.7	3.3	2.6	2.1	13.3
2.1	4.3	2.8	14.8	2.1	11.9	5.9	5.2	9.6	4.1
5.9	4.3	5	10.7	4.1	9.3	4	7.2	4	4.6
12.9	2.2	9.5	10.3	4.9	3.6	10.8	3.5	10.6	2.7
3.5	8.6	4.3	4.6	7.7	2.6	7.1	5.4	5.8	8
2.6	10.8	8.9	3.9	6.3	2.1	2.9	5.8	7.4	4.2
8.6	2.9	7.2	13.3	2.9	9	6.2	11.7	10.3	9
2.7	5.3	5.6	2.2	2.6	11	10.3	5.6	6.3	7.7

1.

Save this data set to your computer or flash drive. Create a histogram of this data and describe the histogram (number of classes, class interval, comment about shape.) What did the histogram reveal about the data that you found of most interest?

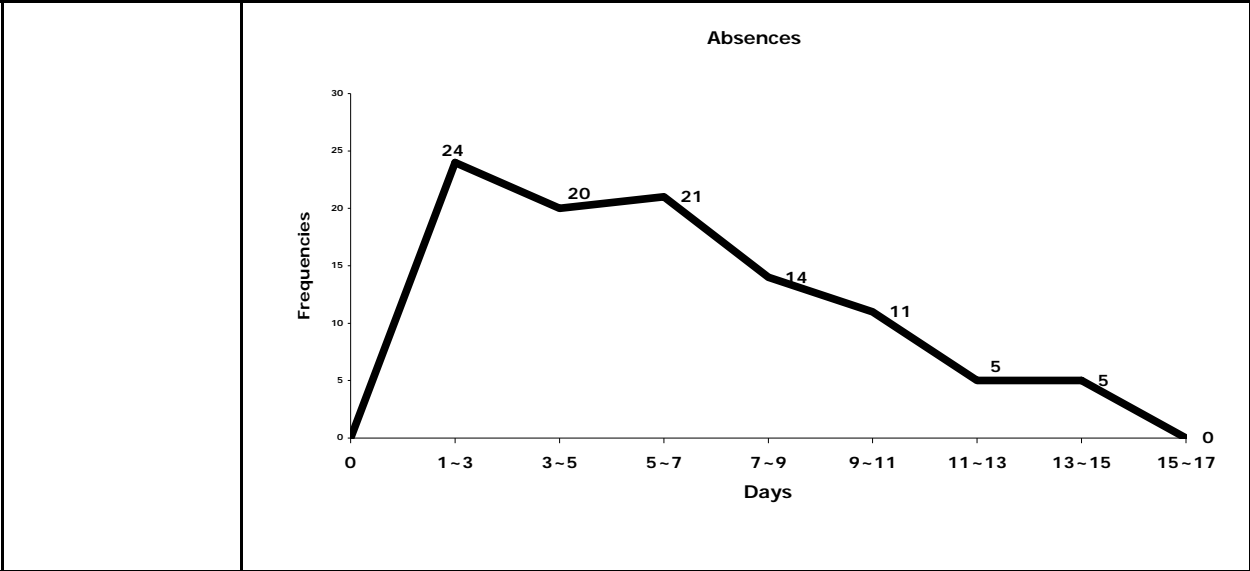


Answer

The histogram I created had seven columns with a width of two days. The first class was 1 to 3 days and the last 13 to 15 days. On the left the graph is quite uniform through 7 days, then frequencies start to drop. The rightmost two columns have identical frequencies describing a long tail. Almost $\frac{2}{3}$ ^{ths} of the individuals averaged less than 7 days absent per year.

2.

Using the same data set as in the previous question, create a frequency polygon for this data and describe it. How does your description differ from that of the histogram in the previous question? What did you learn that was new about the data set by creating the frequency polygon?



Answer

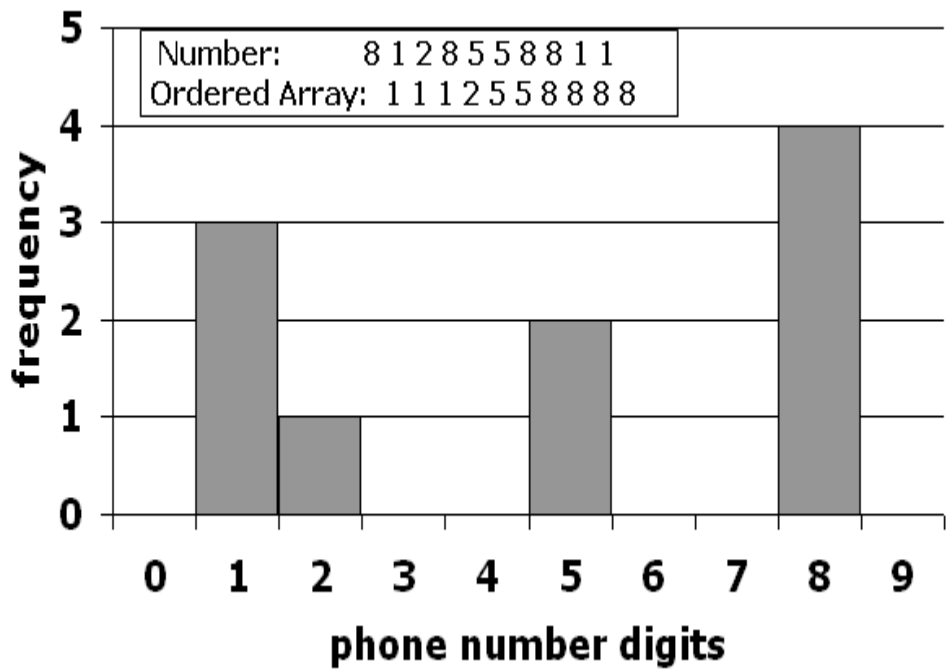
The polygon has the same number of segments as the histogram had columns however, forcing the line graph to cross the X axis at the left and right extremes gives an impression of continuity that the separate bins in the histogram represses. The primary difference in the two graphs is the difference between the line and the columns. The polygon appears to emphasize the change in frequency from one bin to the next more than did the histogram due to the ease of comparing slopes of line segments. Once the bulk of the columns was removed, it became clear that the class with the highest frequency was that of 1 to 3 days. While this was certainly visible in the histogram, the polygon revealed it quite clearly. Additionally, the combined line segments from 7 days to 11 days seems to suggest that the frequencies decline at virtually the same rate.

3.

Using the same data set as in the previous questions, create a relative frequency ogive and describe it. How does your description differ from the two previous graphs you created? What did you learn about the data that you did not see in the previous two graphs?

	<p style="text-align: center;">Relative Frequency Ogive</p> <table border="1" style="margin-left: auto; margin-right: auto;"> <thead> <tr> <th>Days</th> <th>Relative Frequency</th> </tr> </thead> <tbody> <tr><td>0</td><td>0.00</td></tr> <tr><td>3</td><td>0.24</td></tr> <tr><td>5</td><td>0.44</td></tr> <tr><td>7</td><td>0.65</td></tr> <tr><td>9</td><td>0.79</td></tr> <tr><td>11</td><td>0.90</td></tr> <tr><td>13</td><td>0.95</td></tr> <tr><td>15</td><td>1.00</td></tr> </tbody> </table>	Days	Relative Frequency	0	0.00	3	0.24	5	0.44	7	0.65	9	0.79	11	0.90	13	0.95	15	1.00
Days	Relative Frequency																		
0	0.00																		
3	0.24																		
5	0.44																		
7	0.65																		
9	0.79																		
11	0.90																		
13	0.95																		
15	1.00																		
Answer	<p>The ogive describes a gentle upward sloping curve which seems to suggest little difference in the rates of change from one category to the next. In fact, the slope of each segment from 3 days to 7 days is virtually the same as is the slope of each segment from 11 to 15. This curve is remarkably close to a straight line. It appears to muffle the actual changes in frequency from one category to the next particularly when compared to the histogram. This graph reveals very clearly the 65% of observations that are less than 7 days on average. However, it subdues the fact that the first category has the highest frequency so clearly visible in the polygon. The graph does reveal that the bulk of the data is in the lower categories and the decreasing slope reveals the long right tail previously seen in the data.</p>																		
4.	<p>On a piece of paper, write down the 10 digits of your phone number, including area code. Arrange them in an ordered array, then sketch a histogram of the numbers, one digit to a bin. Here is an example using my office phone number.</p>																		





Draw an arrow under the horizontal axis at the location of the mean. In the answer area below, describe the picture you drew and explain why the mean is the "balance point" of the data set. What is being balanced?

<p>Answer</p>	<p>The ordered array of my phone number is 1 1 1 2 5 5 8 8 8 8 . The histogram is pretty scanty, with a cluster of two columns at 1 and 2, then a column at 5 and the highest column at 8, so lots of gaps. If I had to describe its shape, I would say it makes a U. The mean is at 4.7, which is a little closer to the column at 8 than the column at 1. The deviation between the mean and the minimum value is $4.7 - 1 = 3.7$. For the maximum value it is $8 - 4.7 = 3.3$. There are 4 8s in the data set, and they are balancing the 3 1s and the 2, yet the minimum is a little farther away than the maximum. Ah! The number of values on either side of the mean is not what is being balanced, but the distances between the values and the mean on either side of it. Deviations of values less than the mean sum to -13.8 and those of values greater than the mean sum to $+13.8$.</p>
<p>5.</p>	<p>Consider the histogram you sketched for the previous question. Change the vertical axis to be relative frequencies. Now stack each bin on top of the first one. (You want to end up with one column that is the sum of all the columns in height.) How tall is the column in terms of relative frequencies? Why? Refer to your drawing when you answer.</p>
<p>Answer</p>	<p>The frequencies from before were 3, 1, 2, 4 and as relative frequencies they are 0.3, 0.1, 0.2, and 0.4. When I stack all the columns on top of one another I get one very tall column at the value of 1 on the X axis and the total relative frequency is the sum of the relative frequencies, or 1 on the Y axis. The whole equals the sum of the parts.</p>

6.	Back at your original histogram from question 4, mark under the horizontal axis the median of the data set. Sketch the "deviations" of the observations from the median. How do these "deviations" compare to the deviations of the observations from the mean? Is the median a balance point of any kind?
Answer	The median is 5. These "deviations" are not canceling as they did in the first question with the mean, in fact I end up with a total deviation of (-3). The 2 5s add nothing to the sum (since 5-5=0), and that seems to matter. The median, however, does balance the NUMBER of observations above and below it, by definition. Again, ignoring the 5s, there are 4 numbers above and 4 numbers below the median.
7.	Describe what this formula: $\sigma^2 = \frac{\sum_i (X_i - \mu_x)^2}{N}$ is telling you to do. Do NOT tell me what the result is or what it is a formula for. Describe the actions it is directing you to take.
Answer	The formula is saying to first calculate the difference between each observation in the data set and the mean of the data set. Each difference should then be squared and then the squared values are added together. The resulting sum is divided by the number of observations in the data set.