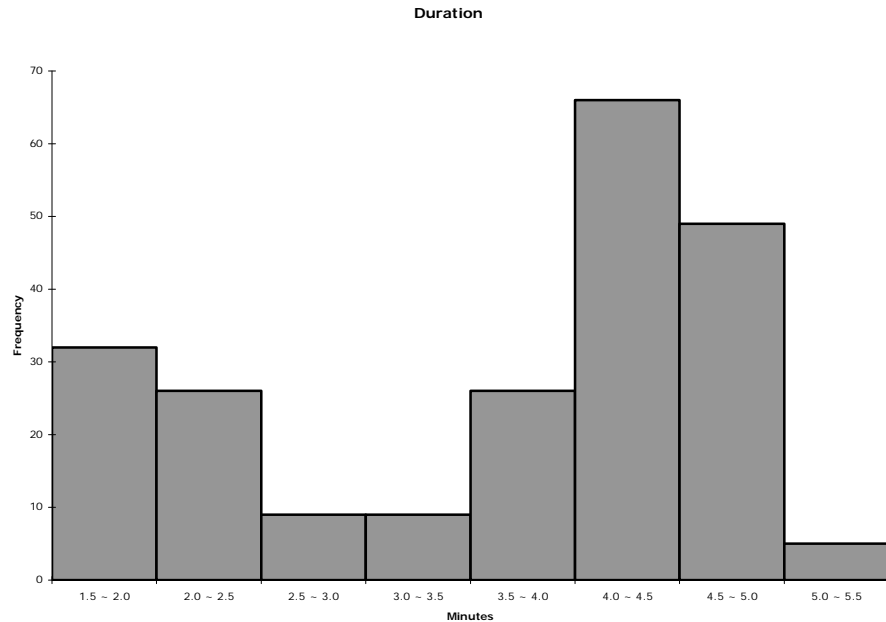


Suggested Answers for WarmUps for Lesson 08

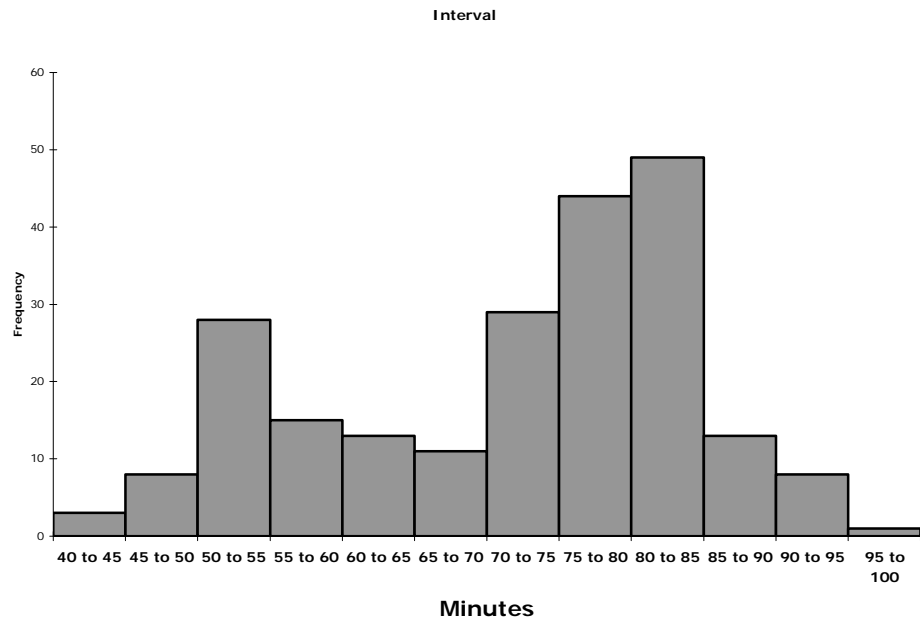
1.

Thoroughly analyze and describe the distributions of the two variables Interval and Duration and the relationship between them. In your response state specifically what methods you used and in what order, as well as your descriptions and conclusions.

Answer

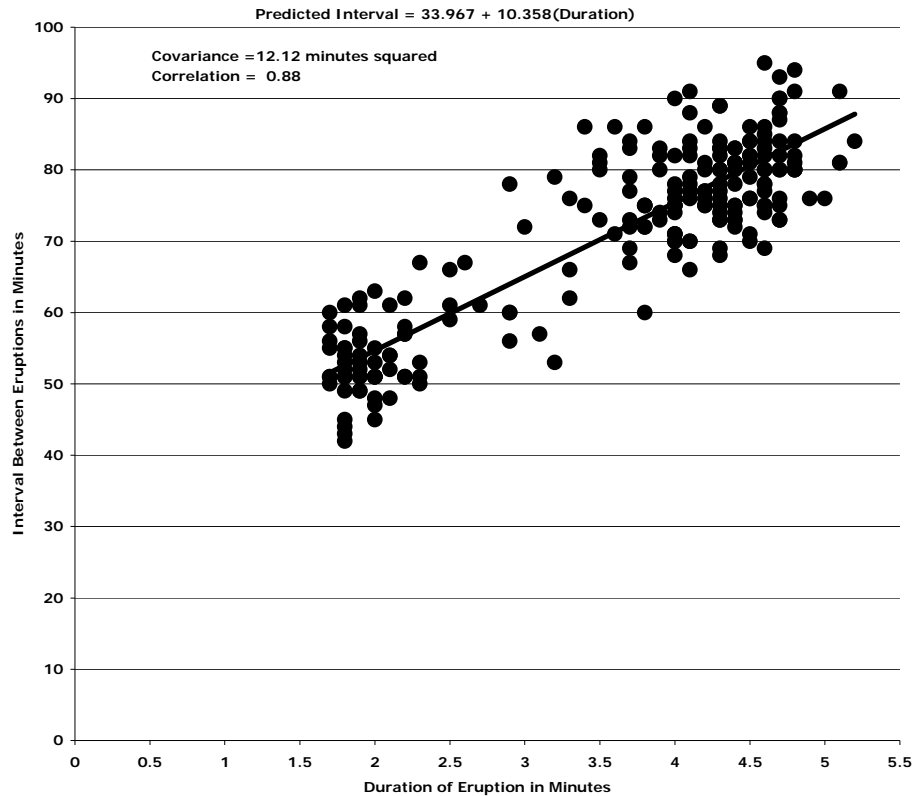


The variable Duration has an interesting bi-modal distribution as can be seen above. While the modes are not identical in height, they are both clearly “local” modes. The bi-modal nature of the distribution makes mean and median values somewhat valueless, but nonetheless, they are, respectively, 3.58 and 4 minutes. Note that while the median is a member of the modal class of 4 to 4.5 minutes, the mean is not, but instead is a member of the next class lower, 3.5 to 4 minutes. Half of the observations are between 4 and 5.5 minutes, but the lower half are between 1.5 and 4 minutes, a range of 1.5 minutes compared to that of 2.5 minutes. There is more variation among the shorter durations than among the longer durations. Equally the standard deviation of 1.08 minutes is only somewhat valuable. Of more interest is the standard deviation of the lower 50% of observations, 0.86 minutes and that of the upper 50% of observations 0.27 minutes, confirming the conclusion drawn based on the ranges of the two portions of the distribution. Overall a visitor to the park could reasonably expect an eruption to last 4 to 4.5 minutes, but a significantly large number of the eruptions (almost 50%) are less than 4 minutes and can be as short as 1.5 minutes.



The variable Interval also exhibits a bimodal distribution quite similar to that of Duration. Following the method used to describe Duration, Interval's mean and median are 71 and 75 minutes respectively; both mean and median are in the same class, which is NOT either of the modal classes. The upper half of the observations are between 75 and 100 minutes, while the lower half are between 40 and 75 minutes, a range of 25 minutes and 35 minutes respectively. The variability of shorter Intervals is greater than that of the longer intervals, as was the case with Duration. The standard deviation of the data set is 12.8 minutes, but of more interest is the standard deviation of the lower and upper 50% of observations, 9.6 and 4.8 minutes respectively. The dispersion of the lower half is twice that of the upper, confirming the evidence of the ranges. Overall a visitor to the park could typically expect an eruption to occur between 76 and 86 minutes apart, although half of the eruptions occur after a shorter interval, some of which can be as short as 42 minutes.

The analysis of the relationship between the variables in this data set involved the use of a scatter plot, the calculation of a covariance, a correlation coefficient, and a least squares line. A scatter plot of the Old Faithful data is seen below.



Included in the plot are the results of a covariance calculation, a correlation calculation and a least squares line, the formula of the line superimposed on the data points in the scatter plot. The scatter plot shows a strong linear relationship between the duration of an eruption and the interval until the next eruption. Additionally, there are two areas where data is strongly clustered. I suspect that these correspond to the two “modes” that were visible in both of the variables when looked at alone. In general, it appears that short eruptions and brief intervals go together, as do long eruptions and longer intervals.

The Covariance is uninformative on its own, but when standardized into a correlation, the perceived linear relationship is relatively strong and positive, 0.88. The two clusters of data, as did the two modes in the individual series suggest that it might be appropriate to divide the data set for further insight, however that is not necessary at this point.

Some checking of various geologic sources has revealed that in geysers the duration of an eruption is believed to be causal to the length of the interval until the next eruption. Short eruptions expel relatively small amounts of water and steam, and it thus takes less time for the geyser to build pressure back up to the eruption point. Likewise, long eruptions expel more water and steam, and consequently it takes a longer time for the pressure to reach the eruption point again.

This is sufficient to allow the analysis to continue with calculating a least squares line, since causation is established. The resulting least squares line suggests that

for an eruption of no duration (well outside the range of the data set, and hence not informative) the next eruption would occur in about 34 minutes. This is quite contrary to theory but apparently necessary to the accurate prediction of an interval for a duration that is within the data set. The equation further suggests that for each additional minute of the duration of an eruption, the interval until the next eruption increases by 10.4 minutes. The equation provides sufficient information to attempt a prediction of an interval between eruptions. It would be of interest to see how well it does with such predictions. One remaining mystery is the two clusters of data points so strikingly revealed. Analysis using time series might reveal additional information that would explain this phenomenon.

Consider the contingency table found here as you answer the remaining questions.

The information in the table below is in the form of relative frequencies. It pertains to the great ocean liner Titanic which sank on her maiden voyage in April of 1912. The values are the frequencies of passengers and crew sorted by Passenger Class and Survival.

		Passenger Class				
		Crew	First	Second	Third	Total
Survival	No	0.096	0.055	0.076	0.24	0.468
	Yes	0.306	0.092	0.054	0.081	0.532

2. Describe completely the distribution of the variable Passenger Class.

Answer The bulk of the passengers on the Titanic were crew members (40%) followed by Third Class passengers (32%.) First and Second Class passengers composed the remaining 28% in very close proportions (15% and 13%).

3. Describe completely the distribution of the variable Survived.

Answer While the majority of passengers on the Titanic survived, 53%, it is a very narrow majority, and it also means that 47% of passengers died.

4. Provide evidence of the dependence or independence of Passenger Class and Survived. Explain your work

		Passenger Class					Original Relative Frequencies
		Crew	First	Second	Third	Total	
Survival	No	0.096	0.055	0.076	0.24	0.468	
	Yes	0.306	0.092	0.054	0.081	0.532	
Total		0.402	0.148	0.129	0.321	1	
		Products of Marginals					
		Crew	First	Second	Third		
No		0.188	0.069	0.060	0.150		
Yes		0.214	0.079	0.069	0.171		

	Crew	First	Second	Third	Difference: Original-product of marginals.
No	-0.092	-0.014	0.016	0.090	
Yes	0.092	0.013	-0.015	-0.090	

The table above has the original relative frequencies in the first tier. The second tier contains the product of the marginal frequencies for each interior cell. For example, 0.188 in the intersection of "Crew" and "No" is the product of the probability of being a crew member (0.402) and the probability of not surviving (0.468). If Passenger Class and Survived are independent, the expected probability in the intersection of "Crew" and "No" is 0.188, or about 19%. However the observed relative frequency is 0.096, about 10%, in that cell. Only about half as many crew members died as would have been expected. (The difference between actual (top tier) and expected (middle tier) is recorded in the bottom tier.) These calculations are repeated for all joint probability cells.

Those cells which have major differences between observed and expected frequencies have been highlighted in yellow. Half as many crew **died** than expected, and half as many Third Class passengers **survived** than expected. None of the expected frequencies are identical to the observed, but most of the difference lies in Crew and Third Class passengers. This is strong evidence that the class of passenger and whether that passenger survived are dependant in some way.

5.	Which was more likely, given that one did not survive, that person was a crew member, or, given that one was a third class passenger, that person survived. Explain how you drew your conclusion and include any calculations you made.
Answer	$P(\text{Crew} \text{No}) = 0.096/0.468 = 0.205128$ $P(\text{Yes} \text{Third}) = 0.081/0.321 = 0.252336$ It was more likely that a person survived given he or she was a third class passenger than it was that a person was a member of the crew given he or she did not survive.