Eruptions of the “Old Faithful” geyser


Data Files: geyser1.dat, geyser2.dat

A geyser is a hot spring that occasionally becomes unstable and erupts hot water and steam into the air. The “Old Faithful” geyser at Yellowstone National Park in Wyoming is probably the most famous geyser in the world. Visitors to the park try to arrive at the geyser site to see it erupt without having to wait too long; the name of this geyser comes from the fact that eruptions follow a relatively stable pattern. The National Park Service erects a sign at the geyser site predicting when the next eruption will occur. Thus, it is of interest to understand and predict the interval time until the next eruption.

The following analysis is based on a sample of 222 intereruption times taken during August 1978 and August 1979 (data source: Applied Linear Regression, 2nd. ed., by S. Weisberg). The first step in any data analysis is simply to look at the data. A histogram gives a good deal of information about the distribution of intereruption times, suggesting some interesting structure. Interval times are in the general range of 40 to 100 minutes, but there are apparently two subgroups in the data, centered at roughly 55 minutes and 80 minutes, respectively, with a gap in the middle.

A stem–and–leaf display gives a similar impression:
Eruptions of the “Old Faithful” geyser

STEM AND LEAF PLOT OF INTERVAL Intereruption time (minutes)

LEAF DIGIT UNIT = 1
1 2 REPRESENTS 12.

STEM LEAVES
3 4* 234
11 4. 55788999
39 5* 0011111111111122233334444
54 5. 5555666777788889
67 6* 000011112223
78 6. 66677788999
107 7* 00000111112222333333344444
(44) 7. 5555555555555566666666666667777777777888888889999
71 8* 00000000000011111111112222222222333333333333444444444
22 8. 566666678889
9 9* 00011134
1 9. 5

222 CASES INCLUDED 0 MISSING CASES

The stem–and–leaf display can give additional information as well. For example, the first column of the plot above is a cumulative frequency column that starts at both ends and meets in the middle. The row that contains the median is marked with parentheses around the count of observations for that row.

Not all exploratory techniques are as effective for this type of data. A boxplot (sometimes called a box–and–whisker plot) of the interval times shows that interval times are in the general range of 40 to 100 minutes, but the bimodal distribution is hidden by the form of the plot:

Box and Whisker Plot

Summary statistics, as given below, look informative, but that is somewhat misleading. For example, the mean intereruption time of about 71 minutes doesn’t actually describe a typical result from either subgroup. A useful rule–of–thumb is that
roughly 95% of the observations will lie within two standard deviations of the mean, or (here) \( 71 \pm 25.6 = (45.4, 96.6) \). In fact, all but five of the 222 observations fall in this interval, which is more than we would expect. It is important to remember that summary measures are only trustworthy when data values come from a homogeneous population, which is not the case here.

DESCRIPTIVE STATISTICS

<table>
<thead>
<tr>
<th>INTERVAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
</tr>
<tr>
<td>MEAN</td>
</tr>
<tr>
<td>SD</td>
</tr>
<tr>
<td>MINIMUM</td>
</tr>
<tr>
<td>MEDIAN</td>
</tr>
<tr>
<td>MAXIMUM</td>
</tr>
</tbody>
</table>

How, then, can we help the tourists? We need more information. One readily available characteristic of the geyser is the duration of the previous eruption. We could think of our data as pairs of the form (duration of eruption, time until next eruption). Here is a scatter plot of those pairs:

We notice two dominant effects here: there are indeed two distinct subgroups, and a longer eruption tends to be followed by a longer time interval until the next eruption. The existence of two subgroups in this type of data is rare, but not unheard of; J.S. Rinehart, in a 1969 paper in the *Journal of Geophysical Research*, provides a mechanism for this pattern based on the temperature level of the water at the bottom of a geyser tube at the time the water at the top reaches the boiling temperature. That a shorter eruption would be followed by a shorter intereruption time (and a longer eruption would be followed by a longer intereruption time) is also consistent with Rinehart’s model, since a short eruption is characterized by having more water at the bottom of
the geyser being heated short of boiling temperature, and left in the tube. This water has been heated somewhat, however, so it takes less time for the next eruption to occur.

A long eruption results in the tube being emptied, so the water must be heated from a colder temperature, which takes longer. A. Azzalini and A.W. Bowman provide further discussion of statistical analysis based on this model in a 1990 paper in *Applied Statistics*.

The properties of these two kinds of eruptions can be examined in more detail by separating the observations based on duration of the previous eruption. The following summary statistics are for intereruption times based on separating the eruptions by whether the duration is less than, or greater than, three minutes:

### DESCRIPTIVE STATISTICS

<table>
<thead>
<tr>
<th>INTERVAL</th>
<th>FOR ERUPTION DURATION &lt; 3 MINUTES</th>
<th>FOR ERUPTION DURATION &gt; 3 MINUTES</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>N</strong></td>
<td>67</td>
<td>155</td>
</tr>
<tr>
<td><strong>MEAN</strong></td>
<td>54.463</td>
<td>78.161</td>
</tr>
<tr>
<td><strong>SD</strong></td>
<td>6.2989</td>
<td>6.8911</td>
</tr>
<tr>
<td><strong>MINIMUM</strong></td>
<td>42.000</td>
<td>53.000</td>
</tr>
<tr>
<td><strong>MEDIAN</strong></td>
<td>53.000</td>
<td>78.000</td>
</tr>
<tr>
<td><strong>MAXIMUM</strong></td>
<td>78.000</td>
<td>95.000</td>
</tr>
</tbody>
</table>

Boxplots also can be stacked side–by–side on the same plot, separated by subgroups, to provide a graphical representation of this pattern:

Based on these statistics and plots, a simple prediction rule would be that an eruption of duration less than 3 minutes will be followed by an intereruption interval of about 55 minutes, while an eruption of duration greater than 3 minutes will be followed by an intereruption interval of about 80 minutes. Further, the latter longer intereruption time would be expected to occur about two–thirds of the time.
We can evaluate the effectiveness of this prediction rule by testing it on new data. Azzalini and Bowman, in their 1990 *Applied Statistics* article, gave interruption and duration times for 296 eruptions in August 1985, which can be used for this purpose (66 of the eruption durations occurred at night, and were recorded as simply “Short” or “Long,” which we will treat as less than or greater than 3 minutes, respectively; 2 durations were listed as “Medium,” which will be ignored here). Boxplots of the 1985 eruption durations show that apparently the relationship between eruption duration and the previous interruption time had not changed appreciably in the six intervening years (which is what we would expect), which suggests that the prediction rule will work well:

![Box and Whisker Plot](geyser5.ps)

Duration time short (< 3 minutes) or long (> 3 minutes)

296 cases   2 missing cases

We can examine the error in predicting interruption time from the simple prediction rule using August 1985 data. For example, the first eruption given lasted 4.0 minutes, and was followed by a 71 minute interruption time. Since the duration was greater than 3 minutes, we would have predicted the interruption time to be 80 minutes. Thus, the error made was $71 - 80 = -9$ minutes. This operation is then repeated for all available eruptions. A histogram of the error made in using the prediction rule shows that, except for three unusual eruptions, the rule predicts the interruption time to within roughly ±20 minutes (and usually to within ±10 minutes).
This performance can be compared with making a prediction based just on the distribution of the intereruption times. Recall that the overall median intereruption time for the 1978–1979 data was 75 minutes; a histogram of the errors from using that as a prediction rule for the 1985 data shows that the errors are bimodal, and dramatically larger than if the previous duration is used, even being off by as much as 30 minutes:

Thus, by using a very simple rule, the National Park Service can try to ensure that visitors to Yellowstone Park will get to see an “Old Faithful” eruption without waiting
too long.

**Summary**

Intereruption times of the “Old Faithful” geyser in Yellowstone National Park are apparently bimodal, with modes centered at around 55 minutes and 75–80 minutes, respectively. These times are directly related to the duration of the previous eruption, with longer eruptions followed by longer intereruption times (and shorter eruptions followed by shorter intereruption times). This pattern is consistent with previously described physical models of geyser eruption. A simple rule that can be used to predict intereruption time is to predict one or the other of the two modal values, based on whether the previous duration was less than, or greater than, three minutes. This rule can be shown to work well on new eruptions of the geyser, supporting its general use.

**Technical terms**

**Boxplot**: a graphical device in which a rectangle is used to summarize the distribution of a batch of data. The top and the bottom of the rectangle represent the third and first quartiles, respectively. The line inside the rectangle represents the median. The lines extending from the top and bottom of the rectangle represent either the actual limits of the data, or the limits of the bulk of the data (with unusual observations being represented by individual symbols [“flagged”] if they are further out). The boxplot is particularly useful for comparing the location and variability of several batches of data, as boxes can be plotted side–by–side on one plot.

**Histogram**: a graphical device to represent the distribution of a batch of data. The data values are usually grouped into mutually exclusive and exhaustive intervals of equal width, and the number of observations in each interval is determined and represented by a vertical bar. In some variations the widths of the intervals are varied, resulting in potentially different appearances in the plot.

**Median**: an estimate of location determined by ordering the observations in the sample, and then choosing the middle one. If the sample size is even, the median is taken to be the average of the middle two observations. The median has the desirable property of not being greatly influenced by an unusual observation (that is, it is a robust estimate).

**Mean**: an estimate of location determined by adding all of the observations and dividing by the number of observations. It is also known as the average, and is certainly the most commonly used location estimate. An undesirable property is that it can be heavily influenced by even one unusual observation. The usual notation for the sample mean is the use of an overbar, as in $\overline{X}$.

**Mode**: a value, or range of values, characterized by having a greater likelihood of occurrence than values around it. A distribution might have one mode (being unimodal), two modes (bimodal), and so on. Another use of the term is as follows: the sample mode is the value that occurs most often in a sample.

**Quartile**: a value corresponding to a “25% point” in a sample. The first quartile is, roughly speaking, the value such that one–quarter of the observations are less than it, while three–quarters of the observations are greater than it. The third quartile is defined similarly, reversing the quantities “one–quarter” and “three–quarters.” The
second quartile corresponds to the median. Different statistical packages often use slightly different definitions of the quartiles for a given sample size. A measure of the variability in a set of data that is insensitive to a few unusual values is the difference between the third and first quartiles, called the *inter–quartile range*.

**Scatter plot**: a method that can be used to study the joint variation of two variables graphically. Each observation is represented by a point on the plot, indexed by the values on the axes. Each axis is used for a different variable. Besides showing how (and whether) two variables are related to each other, scatter plots also can indicate the existence of distinct subgroups in the data.

**Standard deviation**: an estimate of scale (or variability) determined as the square root of the sum of squared differences between the value of an observation and the sample mean, divided by one less than the sample size.

**Stem–and–leaf display**: a graphical device to represent the distribution of a batch of data. Very similar to a histogram, it is often accompanied by additional information about the data, such as cumulative frequencies and the position of the median. The plot represents the data values by their numerical values, providing additional information over the histogram, but the grouping intervals are usually chosen based on using round numbers, rather than in an attempt to provide the most effective plot.