*Project 7.5 Preparing for a 100-year flood

This project assumes that you have read Section 7.6.

Suppose we are undertaking a review of the flooding contingency plan for a community on the Snake River, just south of Jackson, Wyoming. To properly prepare for future flooding, we would like to know the river's likely height in a 100-year flood, the height that we should expect to see only once in a century. This 100-year designation is called the flood's *recurrence interval*, the amount of time that typically elapses between two instances of the river reaching that height. Put another way, there is a 1/100, or 1%, chance that a 100-year flood happens in any particular year.

River heights are measured by stream gauges maintained by the U.S. Geological Survey (USGS)[4]. A snippet of the data from the closest Snake River gauge, which can be downloaded from the USGS[5] or the book's website, is shown below.

```
#
# U.S. Geological Survey
# National Water Information System
 ⋮
#
agency_cd ▷site_no ▷peak_dt ▷peak_tm ▷peak_va ▷peak_cd ▷gage_ht ▷...
5s ▷15s ▷10d ▷6s ▷8s ▷27s ▷8s ▷...
USGS ▷13018750 ▷1976-06-04 ▷ ▷15800 ▷6 ▷7.80 ▷...
USGS ▷13018750 ▷1977-06-09 ▷ ▷11000 ▷6 ▷6.42 ▷...
USGS ▷13018750 ▷1978-06-10 ▷ ▷19000 ▷6 ▷8.64 ▷...
 ⋮
USGS ▷13018750 ▷2011-07-01 ▷ ▷19900 ▷6 ▷8.75 ▷...
USGS ▷13018750 ▷2012-06-06 ▷ ▷16500 ▷6 ▷7.87 ▷...
```

The file begins with several comment lines preceded by the hash (`#`) symbol. The next two lines are header rows; the first contains the column names and the second contains codes that describe the content of each column, e.g., 5s represents a string of length 5 and 10d represents a date of length 10. Each column is separated by a tab character, represented above by a right-facing triangle (▷). The header rows are followed by the data, one row per year, representing the peak event of that year. For example, in the first row we have:

- `agency_cd` (agency code) is `USGS`
- `site_no` (site number) is `13018750` (same for all rows)
- `peak_dt` (peak date) is `1976-06-04`
- `peak_tm` (peak time) is omitted
- `peak_va` (peak streamflow) is `15800` cubic feet per second
- `peak_cd` (peak code) is `6` (we will ignore this)

[4] `http://nwis.waterdata.usgs.gov/nwis`

[5] `http://nwis.waterdata.usgs.gov/nwis/peak?site_no=13018750&agency_cd=USGS&format=rdb`

- `gage_ht` (gauge height) is `7.80` feet

So for each year, we essentially have two gauge values: the peak streamflow in cubic feet per second and the maximum gauge height in feet.

If we had 100 years of gauge height data in this file, we could approximate the water level of a 100-year flood with the maximum gauge height value. However, our data set only covers 37 years (1976 to 2012) and, for 7 of those years, the gauge height value is missing. Therefore, we will need to estimate the 100-year flood level from the limited data we are given.

Part 1: Read the data

Write a function

```
readData(filename)
```

that returns lists of the peak streamflow and gauge height data (as floating point numbers) from the Snake River data file above. Your function will need to first read past the comment section and header lines to get to the data. Because we do not know how many comment lines there might be, you will need to use a `while` loop containing a call to the `readline` function to read past the comment lines.

Notice that some rows in the data file are missing gauge height information. If this information is missing for a particular line, use a value of 0 in the list instead.

Your function should return two lists, one containing the peak streamflow rates and one containing the peak gauge heights. A function can return two values by simply separating them with a comma, .e.g.,

```
return flows, heights
```

Then, when calling the function, we need to assign the function call to two variable names to capture these two lists:

```
flows, heights = readData('snake_peak.txt')
```

Part 2: Recurrence intervals

To associate the 100-year recurrence interval with an estimated gauge height, we can associate each of our known gauge heights with a recurrence interval, plot this data, and then use regression to extrapolate out to 100 years. A flood's recurrence interval is computed by dividing $(n + 1)$, where n is the number of years on record, by the rank of the flood. For example, suppose we had only three gauge heights on record. Then the recurrence interval of the maximum (rank 1) height is $(3 + 1)/1 = 4$ years, the recurrence interval of the second largest height is $(3 + 1)/2 = 2$ years, and the recurrence interval of the smallest height is $(3 + 1)/3 = 4/3$ years. (However, these inferences are unlikely to be at all accurate because there is so little data!)

Write a function

```
getRecurrenceIntervals(n)
```

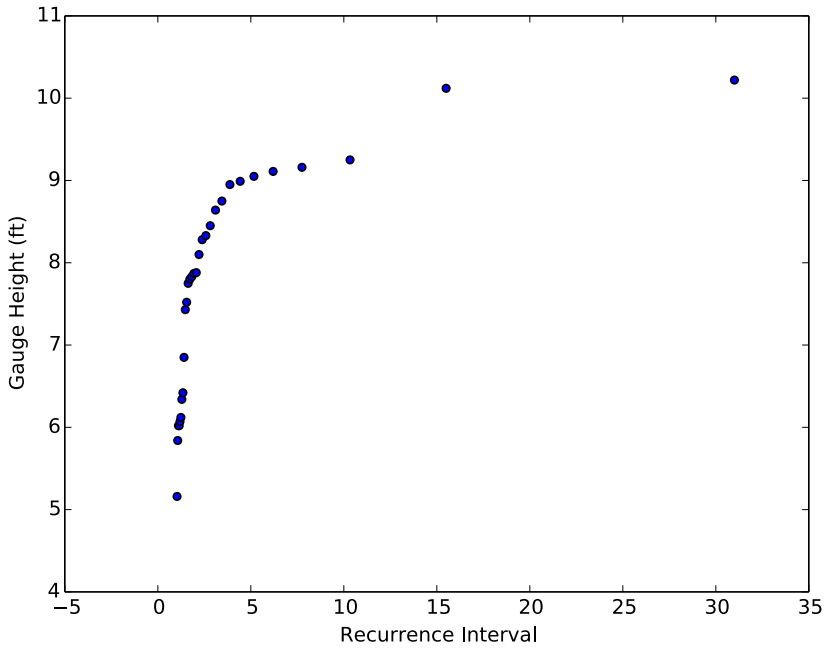

Figure 1 The peak gauge height for each recurrence interval.

that returns a list of recurrence intervals for `n` floods, in order of lowest to highest. For example if `n` is 3, the function should return the list `[1.33, 2.0, 4.0]`.

After you have written this function, write another function

```
plotRecurrenceIntervals(heights)
```

that plots recurrence intervals and corresponding gauge heights (also sorted from smallest to largest). Omit any missing gauge heights (with value zero). Your resulting plot should look like Figure 1.

Part 3: Find the river height in a 100-year flood

To estimate the gauge height corresponding to a 100-year recurrence interval, we need to extend the "shape" of this curve out to 100 years. Mathematically speaking, this means that we need to find a mathematical function that predicts the peak gauge height for each recurrence interval. Once we have this function, we can plug in 100 to find the gauge height for a 100-year flood.

What we need is a regression analysis, as we discussed in Section 7.6. But linear regression only works properly if the data exhibits a linear relationship, i.e., we can draw a straight line that closely approximates the data points.

Question 7.5.1 *Do you think we can use linear regression on the data in Figure 1?*

This data in Figure 1 clearly do not have a linear relationship, so a linear regression

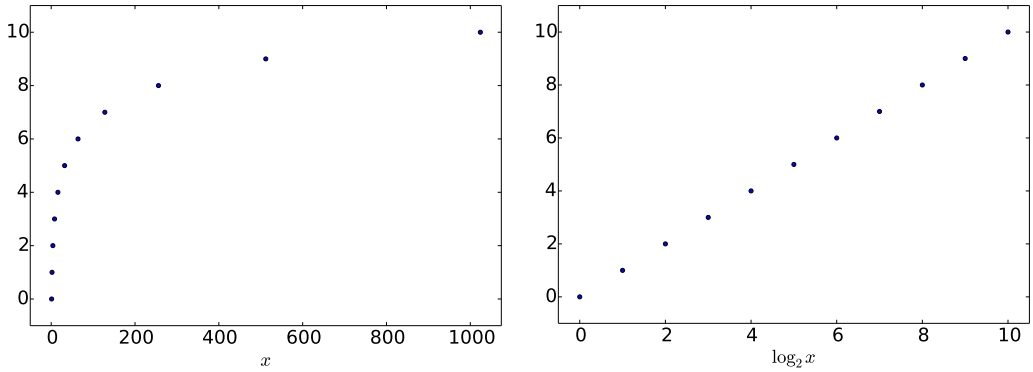

Figure 2 On the left is a plot of the points $(2^0, 0), (2^1, 1), (2^2, 2), \ldots, (2^{10}, 10)$, and on the right is a plot of the points that result from taking the logarithm base 2 of the x coordinate of each of these points.

will not produce a good approximation. The problem is that the x coordinates (recurrence intervals) are increasing multiplicatively rather than additively; the recurrence interval for the flood with rank $r + 1$ is $(r + 1)/r$ times the recurrence interval for the flood with rank r. However, we will share a trick that allows us to use linear regression anyway. To illustrate the trick we can use to turn this non-linear curve into a "more linear" one, consider the plot on the left in Figure 2, representing points $(2^0, 0), (2^1, 1), (2^2, 2), \ldots, (2^{10}, 10)$. Like the plot in Figure 1, the x coordinates are increasing multiplicatively; each x coordinate is twice the one before it. The plot on the right in Figure 2 contains the points that result from taking the logarithm base 2 of each x coordinate $(\log_2 x)$, giving $(0, 0), (1, 1), (2, 2), \ldots, (10, 10)$. Notice that this has turned an exponential plot into a linear one.

We can apply this same technique to the `plotRecurrenceIntervals` function you wrote above to make the curve approximately linear. Write a new function

```
plotLogRecurrenceIntervals(heights)
```

that modifies the `plotRecurrenceIntervals` function so that it makes a new list of logarithmic recurrence intervals, and then computes the linear regression line based on these x values. Use logarithms with base 10 for convenience. Then plot the data and the regression line using the `linearRegression` function from Exercise 7.6.1. To find the 100-year flood gauge height, we want the regression line to extend out to 100 years. Since we are using logarithms with base 10, we want the x coordinates to run from $\log_{10} 1 = 0$ to $\log_{10} 100 = 2$.

Question 7.5.2 *Based on Figure 2, what is the estimated river level for a 100-year flood? How can you find this value exactly in your program? What is the exact value?*

Part 4: An alternative method

As noted earlier in this section, there are seven gauge height values missing from the data file. But all of the peak streamflow values are available. If there is a

linear correlation between peak streamflow and gauge height, then it might be more accurate to find the 100-year peak streamflow value instead, and then use the linear regression between peak streamflow and gauge height to find the gauge height that corresponds to the 100-year peak streamflow value.

First, write a function

```
plotFlowsHeights(flows, heights)
```

that produces a scatter plot with peak streamflow on the x-axis and the same year's gauge height on the y axis. Do not plot data for which the gauge height is missing. Then also plot the least squares linear regression for this data.

Next, write a function

```
plotLogRecurrenceIntervals2(flows)
```

that modifies the `plotLogRecurrenceIntervals` function from Part 3 so that it find the 100-year peak streamflow value instead.

Question 7.5.3 *What is the 100-year peak streamflow rate?*

Once you have found the 100-year peak streamflow rate, use the linear regression formula to find the corresponding 100-year gauge height.

Question 7.5.4 *What is the gauge height that corresponds to the 100-year peak streamflow rate?*

Question 7.5.5 *Compare the two results. Which one do you think is more accurate? Why?*