1 Graphical or Picture Representation of a Distribution

1.1 Ball Bearings

These data are and information can be found in the "Datasets and Stories" article "Modeling the Reliability of Ball Bearings" in the *Journal of Statistics Education* (Caroni 2002, www.amstat.org/publications/jse/v10n3/datasets.caroni.html).

We only consider here the diameter of the ball bearings.

Histograms of the data for the first two companies are given in Figure 1. Relative frequency histograms are shown in Figure 2.

The histogram gives on the y-axis the counts of the number of observations in each bin or interval. The relative frequency histogram gives the proportion of observations in each interval. The shapes of the two types of plots are the same, only the y-axis is different.

From these we can readily see many things about these data. For company 1, the diameters range from 0 to 2 (actually we only know there is one observation in the bin from 0 to $\frac{1}{4}$, and also in the bin from 1.75 to 2). For company 2 these data range from 0 to 1.75, so there is a larger range of bearing sizes for company 1. For company 1, the *central* or *average* (mean) observation is around .5 or .75, and for company 2 the diameter is centred around .5.

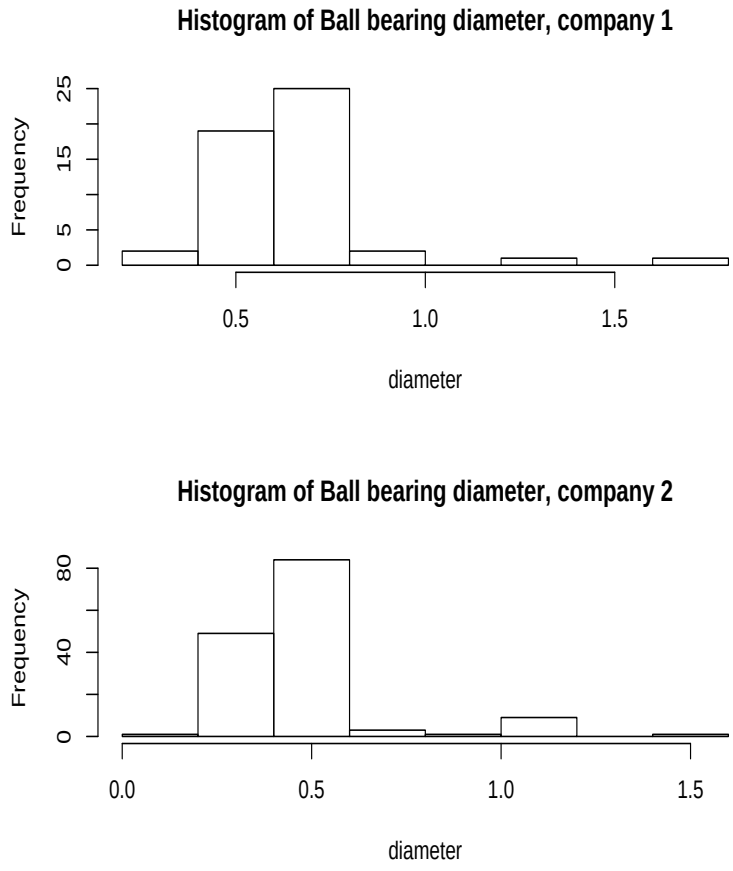

Figure 1: Ball bearing histogram for companies 1 and 2

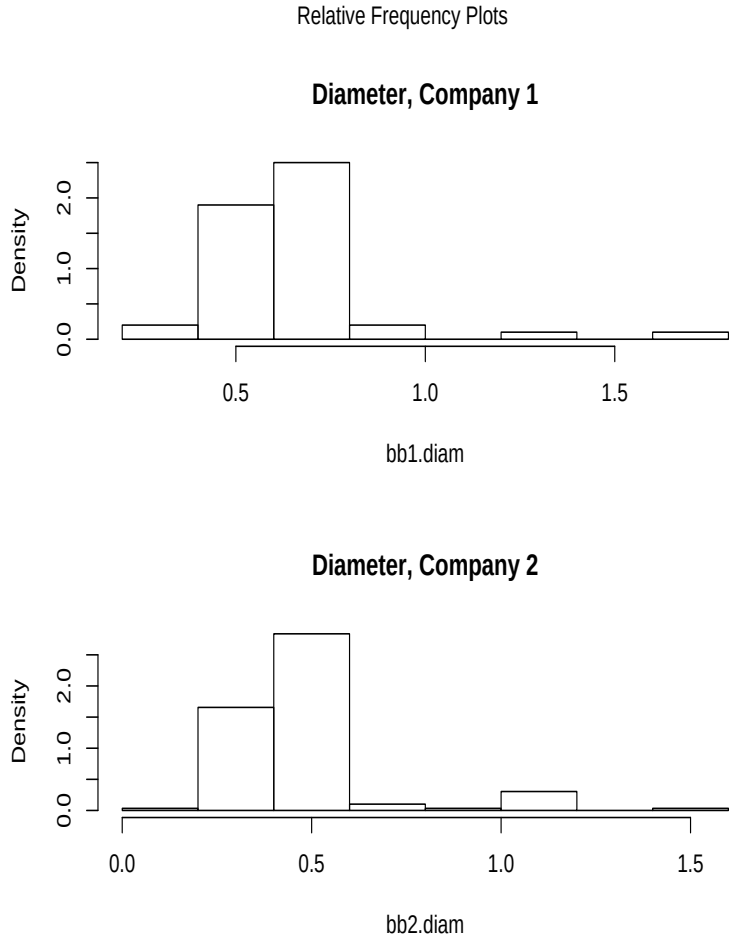

Figure 2: Ball bearing relative frequency histogram for companies 1 and 2

Later we will discuss some numerical summary numbers for distributions, and summary statistics for data. For this purpose we give some summary numbers (statistics) for the ball bearing data in Table 1.

Table 1: Some summary numbers to describe the ball bearing diameter data

Company	mean	standard deviation	variance	median
1	0.631	.225	.0506	.656
2	0.463	.215	.0461	0.469

Box plots are another useful plotting method. It gives means and first and third quartiles. These will be explained later.

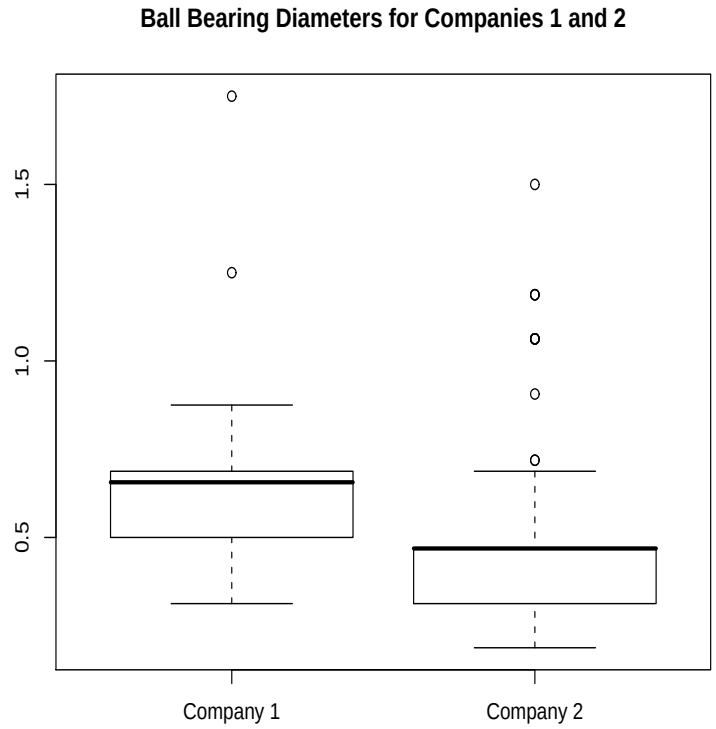

Figure 3: Ball bearing boxplot for companies 1 and 2

1.2 Acid Rain

This data set is the pH level of water samples from Shenandoah National Park, in Virginia state, USA. The data is obtained at the URL

`http://www.cvgs.k12.va.us/DIGSTATS/main/inferant/a_acidrain.htm`

The pH scale falls between 0 and 14 where 0-6.9 is acidic, and 7.1-14 is basic. Pure water has a pH level of 7.0 which is neutral on the scale, but rain water is slightly acidic usually around 5.6 on the scale. Acid rain is defined as having a pH level of 5.6 or lower.

Frequency histogram and relative frequency histogram of the pH reading for a sample of $n = 90$ data points.

See the data file pHlevels.xls

Figures 4 and 5 are the frequency histograms and relative frequency histograms. The frequency histogram counts the number of data values in each interval, while the relative frequency histogram gives the proportion or relative frequency in each interval. The two plots look the same except for the scale on the y-axis.

What causes acid rain and how do we control it? First one has to determine in lakes or other bodies of water are acidic. Such studies lead to the replacement of CFC as the main refridgerant used in air conditioners in cars and homes, and the main coolant used in refridgerators. These replacements occurred in the early 1990s. Acid rain is still a problem but not as frequent

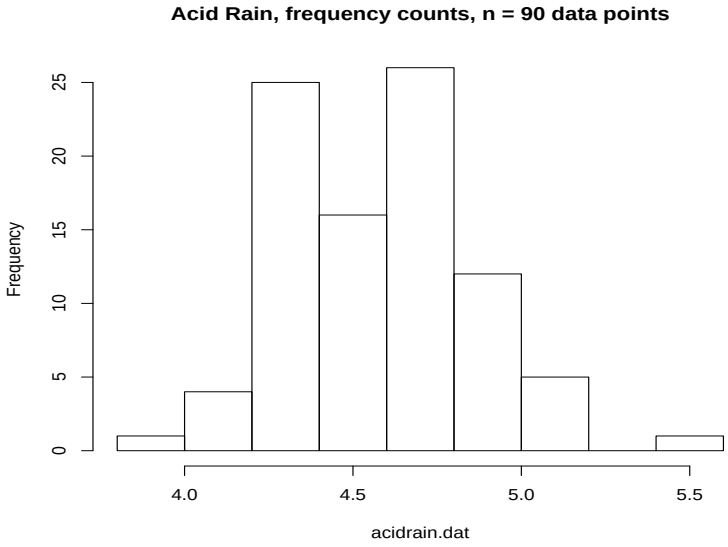

Figure 4: Acid Rain histogram

as in the 1980s. This is only one of many environmental concerns and the use of statistics in that important field.

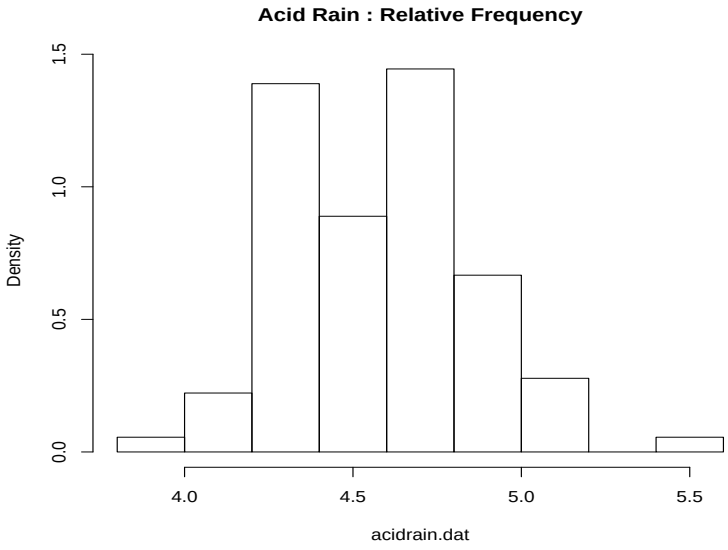

Figure 5: Acid Rain relative frequency histogram

Sometimes data comes in categories, either naturally or by manipulating or summarizing the original data.

Suppose the acid rain was only given by the number of data values in certain intervals, such as

Table 2: Numbers of Observations in Categories

category name	interval	count
strong	< 4.5	45
acidic	4.5 to 5.0	50
weak acidic	> 5.0	6

This type of data may also be as relative proportions instead of counts. Notice that the relative proportions must add to 1, since 100% of the data is in one of the categories. When the relative frequency or proportion is rounded, in this case to 2 digits, these proportions sum to 1.01 (equivalently 101%). This is known as round off error.

Table 3: Numbers of Observations in Categories

category name	proportion
strong	.38
acidic	.56
weak acidic	.07
Total	1.01

This type of categorical data is often plotted as a bar chart or as a pie chart. Suppose we only have the categorical count information. Thus we only know for example there are 6 data points out of the 90 data points that fall

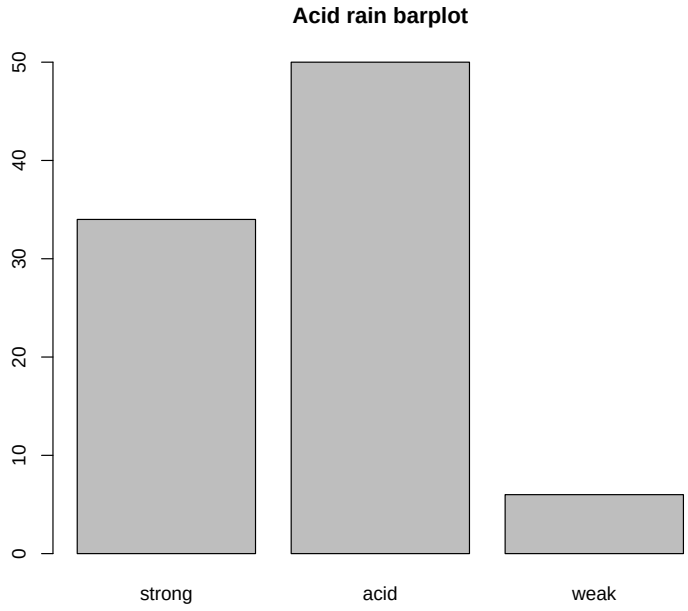

Figure 6: Bar Chart of Acid Rain groups

into category weak acidic, and not the actual values of these 6 data points. For the categorical data from Table 2 we have the bar chart (Figure 6) and pie chart (Figure 7).

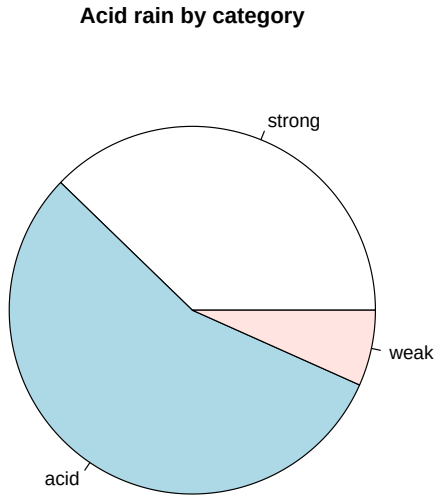

Figure 7: Bar Chart of Acid Rain groups

The acid rain histogram is roughly symmetric and bell shaped. There is a function called the normal distribution (normal density, normal curve, Gaussian distribution, Gaussian curve) that is of this shape.

For reasons beyond this course, this distribution often gives a good description of many natural phenomena or data. It requires two number to specify the distribution. For the acid rain data we have

mean = 4.578	variance = 0.0836	std = $\sqrt{\text{variance}}$ = 0.289

Here std is shorthand for standard deviation.

In Figure 8 we have the relative frequency histogram for the acid rain data with a normal curve overlaid, where the normal curve has mean = 4.57 and standard deviation = 0.28. Notice is one reads areas under the normal curve between two points on the x-axis we get nearly the same number as the area under the histogram. Thus we can think of the normal curve as giving a good *model* to describe the acid rain data, and it requires only two numbers, instead of all 90 needed to specify the histogram.

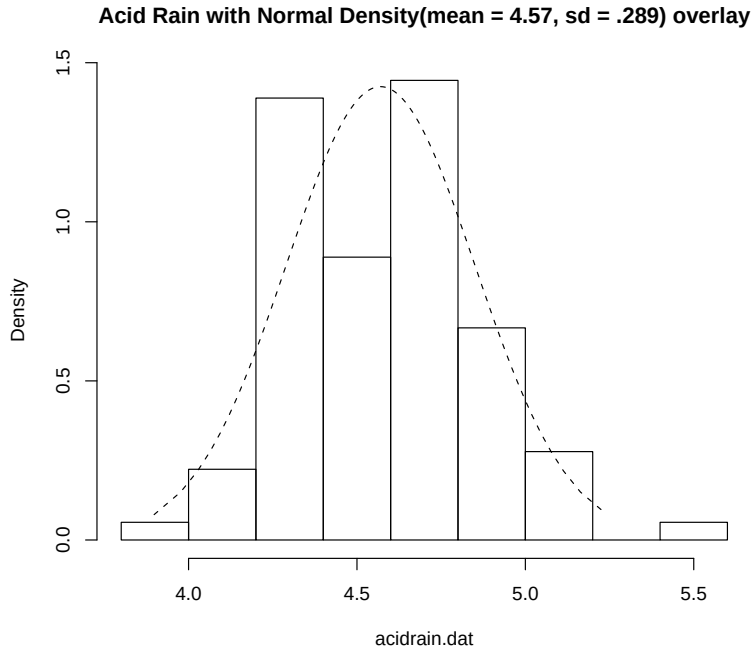

Figure 8: Acid Rain relative frequency histogram and Normal overlay

We do not use this further in SS1024, but those who are interested the normal curve is specified by the function

$$f(x) = \frac{1}{\sqrt{2\pi b^2}} e^{-\frac{(x-a)^2}{2b^2}}$$

where a, b are two numbers. In Figure 8 we use $a = 4.57$ and $b^2 = .0836$ (or equivalently $b = \sqrt{b^2} = 0.289$).

1.3 Fecal Coliform

Fecal coliform is a bacteria that can infect open water. Some of you may have seen references to beaches closed or boil water advisories in this province. These often occur in Ontario, happening several times per year.

From these two histograms it is clear that the coliform counts are quite different when observed on dry days as opposed to wet days. The wet day readings are in a sense higher (larger) than dry days. The fecal content is variable and sometimes one has dry day readings larger than some wet day readings. Nonetheless it is clear that in a *statistical sense* the wet day fecal coliform counts are larger than on dry days.

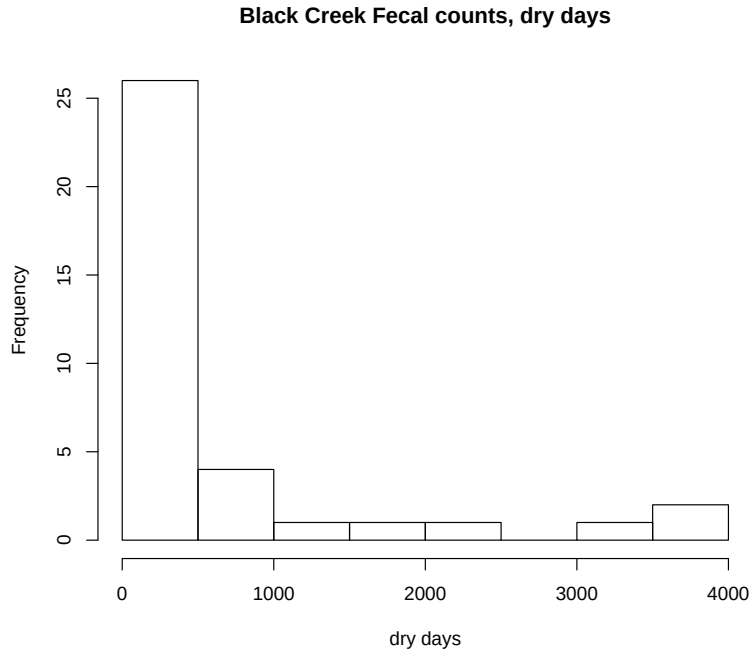

Figure 9: Black Creek fecal coliform on dry days

These shapes cannot be described by the normal curve.

Figures 9 and 10 are examples of skewed distributions, often described as skewed to the right. There are other mathematical models that are used in this case, but these are not discussed in this course. Sometimes these models transform data, for example using the square root of the data or the logarithm of the data.

Some summary data is still useful and is given in Table **??** for the fecal data.

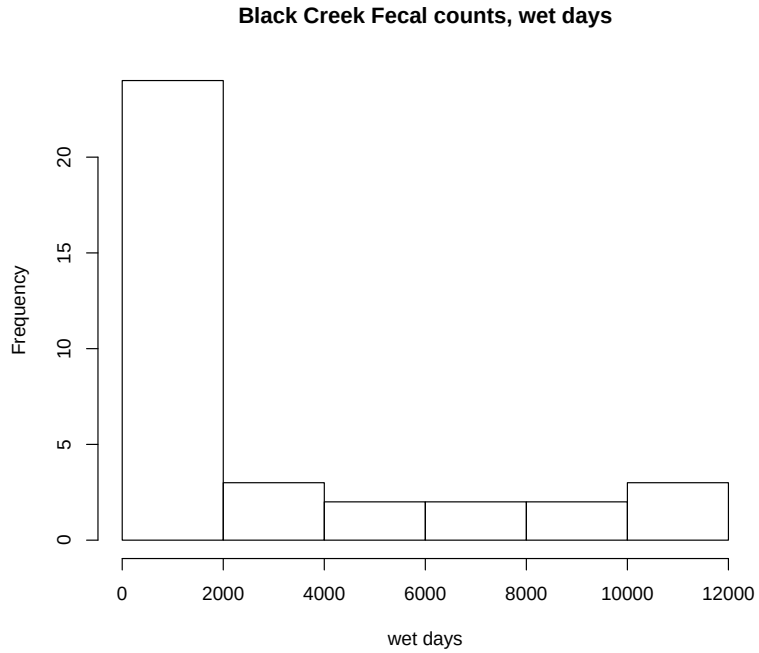

Figure 10: Black Creek fecal coliform on wet days

Financial data, such as stock market data, typically use the natural logarithm of the relative stock price.

Table 4: Summary Statistics for Fecal Data

	Min	1st quartile	Median	mean	3rd quartile	Max
dry	0.0	0.0	100	580.6	625	3800
wet	0.0	400	900	2719	3725	11000