

1 – Descriptive Statistics: Visualizing Data

Some History

The subject of *Statistics* began in the 18th century when governments (i.e., states) in Europe began collecting data about their citizens. A modern definition is:

Statistics is *the branch of mathematics that deals with data*.

The purpose of statistics is to *make sense* of data to support decision-making.

Ronald Fisher (1890-1962) developed foundational concepts and methods of statistical inference in connection with genetics and agricultural data.



Ronald Fisher,
British Statistician

Why Statistics in CST?

The core concepts of statistics and probability are used throughout *data science* (including *machine learning*, *signal processing*, *data mining*) as well as in business, politics, medicine, and in many other areas.

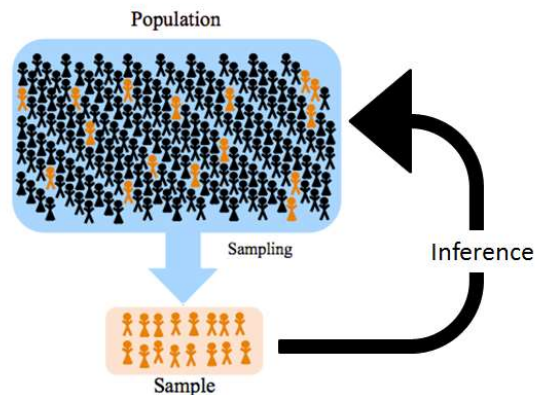
Also, data analytics might be a core feature of your software! For instance, the Whoop app evaluates your sleep and physical recovery by comparing it to your past data.



Parts of Statistics

Descriptive Statistics provides methods for organizing and summarizing large data sets.

Inferential Statistics provides methods for drawing conclusions about a *population* based on data from a *sample*.



Probability Theory is the mathematical link between descriptive and inferential statistics.

Example (Student Survey) Given the *data frame* below (first 10 rows shown), we might try to answer questions like:

- Do most students exercise?
- Are students who smoke more or less likely to exercise?
- What is the typical age of a student?

	Sex	Wr.Hnd	NW.Hnd	W.Hnd	Fold	Pulse	Clap	Exer	Smoke	Height	M.I	Age
1	Female	18.5	18.0	Right	R on L	92	Left	Some	Never	173.00	Metric	18.250
2	Male	19.5	20.5	Left	R on L	104	Left	None	Regul	177.80	Imperial	17.583
3	Male	18.0	13.3	Right	L on R	87	Neither	None	Occas	NA	NA	16.917
4	Male	18.8	18.9	Right	R on L	NA	Neither	None	Never	160.00	Metric	20.333
5	Male	20.0	20.0	Right	Neither	35	Right	Some	Never	165.00	Metric	23.667
6	Female	18.0	17.7	Right	L on R	64	Right	Some	Never	172.72	Imperial	21.000
7	Male	17.7	17.7	Right	L on R	83	Right	Freq	Never	182.88	Imperial	18.833
8	Female	17.0	17.3	Right	R on L	74	Right	Freq	Never	157.00	Metric	35.833
9	Male	20.0	19.5	Right	R on L	72	Right	Some	Never	175.00	Metric	19.000
10	Male	18.5	18.5	Right	R on L	90	Right	Some	Never	167.00	Metric	22.333

1.1 Graphical Methods of Describing Data

Data can be either:

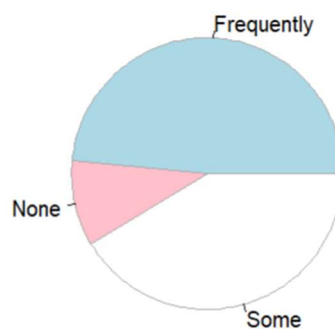
- *non-numerical* (also known as *categorical* or *qualitative*), or
e.g., **Smoke**
- *numerical* (also known as *quantitative*).
e.g., **Age**

Graphs for Non-Numerical Data

Pie Chart

```
> pie(table(survey$Exer))
```

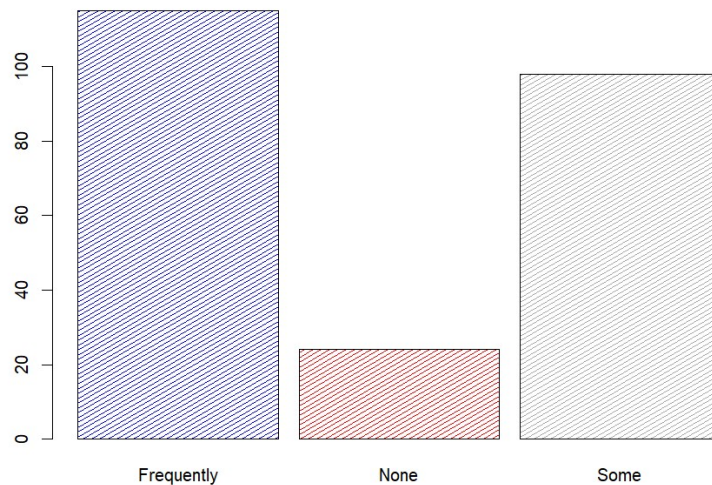
Exercise Habits of Students (n= 237)



Bar Chart

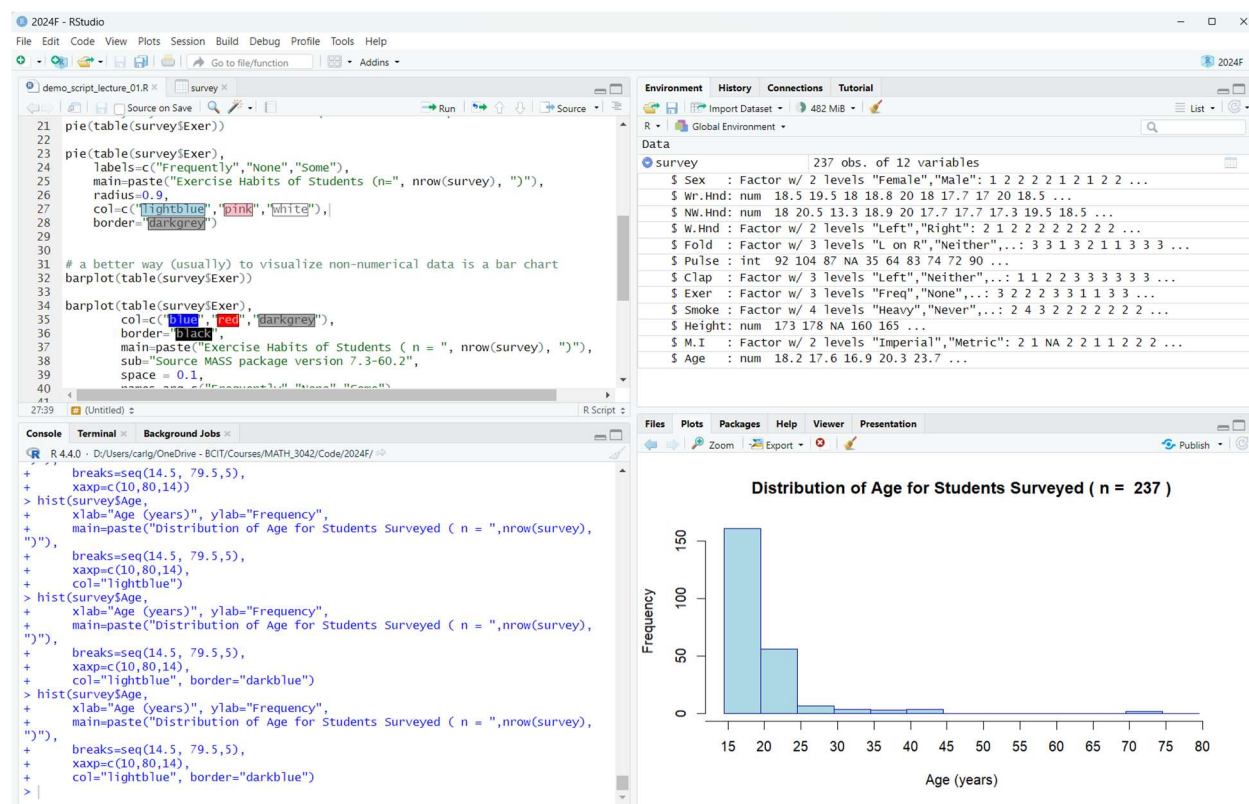
```
> barplot(table(survey$Exer))
```

Exercise Habits of Students (n = 237)



Source MASS package version 7.3-60.2

Sidenote on Software We will be using the programming language R and the coding environment RStudio to create graphs and charts. R is an open-source programming language for statistical computing and graphics. The R language is widely used among statisticians and data miners for developing statistical software and data analysis.



Graphs for Numerical Data

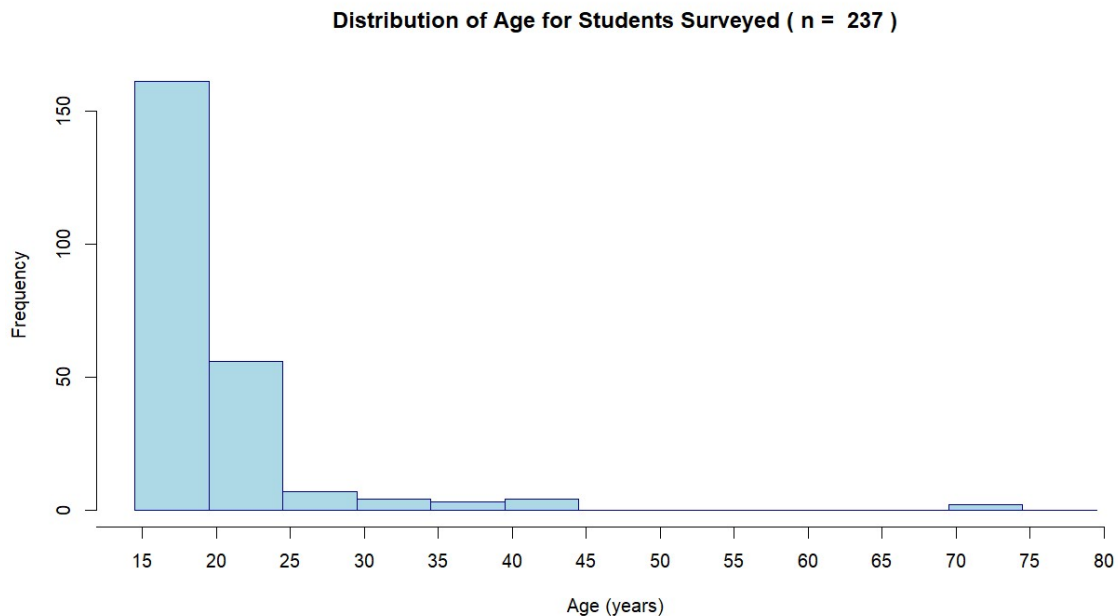
We will spend most of our time in this course dealing with *numerical* variables.

When the set of numerical values is relatively large (> 100 individuals) then it becomes impossible to understand the data without the help of graphs or numerical summaries.

Histograms, a Quick Look

Example (Student Survey) Again, we use the student survey data frame mentioned above. We can visualize the students' ages in several ways. The most important way is using a *histogram*, shown below.

```
> hist(survey$Age)
```



Key Features of a Histogram

- Variable values X go along the horizontal axis.
- The X axis is divided into *classes*.
- The *frequency* of a given class is shown as the *height* of the rectangle for that class.
- Classes have *equal width*.
- There are no *gaps* between classes.
- Each *data value* falls into a class.

Example (Old Faithful) Another built-in data is *faithful*, which contains data describing eruptions of the geyser called “Old Faithful” in Yellowstone, USA.

We can look at the data using the command `View(faithful)`.



	eruptions	waiting
1	3.600	79
2	1.800	54
3	3.333	74
4	2.283	62
5	4.533	85
6	2.883	55
7	4.700	88
8	3.600	85
9	1.950	51
10	4.350	85

In this data set:

- eruptions = the duration (in minutes) of observed eruptions
- waiting = the time (in minutes) in between one eruption and the next eruption

Stem-and-leaf

Let's focus on the *waiting* data. A simple (although old-fashioned) way to visualize the data is a *stem-and-leaf plot*.

```
> stem(faithful$waiting)
```

The decimal point is 1 digit(s) to the right of the |

```

4 | 3
4 | 55566666777788899999
5 | 0000011111122222333333444444444
5 | 555555666677788889999999
6 | 00000022223334444
6 | 555667899
7 | 00001111123333333444444
7 | 55555556666666667777777778888888888888889999999999
8 | 000000001111111111112222222222333333333333333334444444444
8 | 5555556666666778888888999
9 | 00000012334
9 | 6

```

In a stem-and-leaf plot, each data value is cut into a “stem” and a “leaf”. The leaf is typically the final numerical digit. The stem is everything before the leaf.

7 9

If the data values have more than three digits of precision, it may be necessary to round each value to three digits first. For the *eruptions* data, a typical data value is cut like this:

2 . 2 8 3


```
> stem(faithful$eruptions)
```

The decimal point is 1 digit(s) to the left of the |

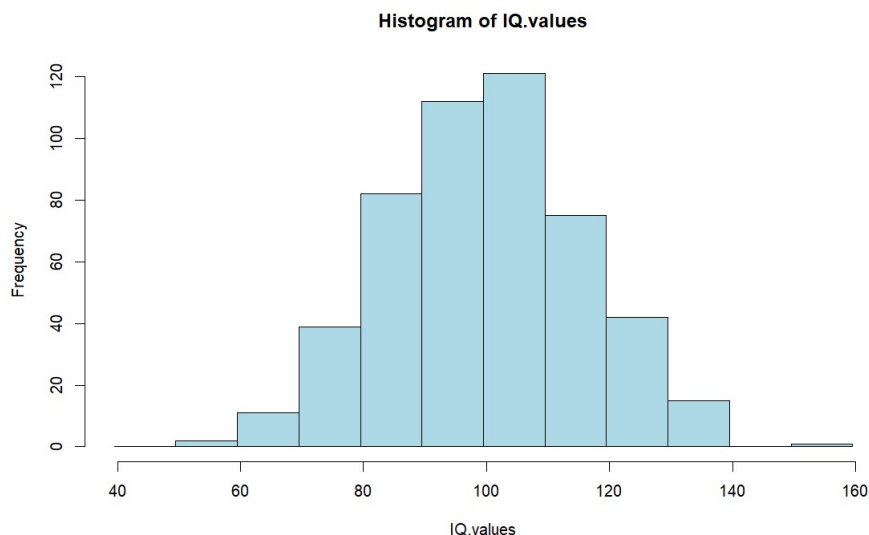
```
16 | 07035555588
18 | 0000222333333557777777888822335777888
20 | 00002223378800035778
22 | 0002335578023578
24 | 00228
26 | 23
28 | 080
30 | 7
32 | 2337
34 | 250077
36 | 0000823577
38 | 2333335582225577
40 | 0000003357788888002233555577778
42 | 03335555778800233333555577778
44 | 0222233555778000000023333357778888
46 | 0000233357700000023578
48 | 00000022335800333
50 | 0370
```

What can we conclude about the duration of *eruptions* at Old Faithful?

Frequency Distributions and Histograms

A stem-and-leaf plot is a crude form of a *histogram*, which we will now cover in more detail.

A histogram is always derived from a *frequency distribution* based on a choice of *classes* for the variable in question.



Example (Old Faithful) One possible choice of classes for the *eruptions* data and the corresponding frequencies is given by the *frequency distribution* shown in the table below.

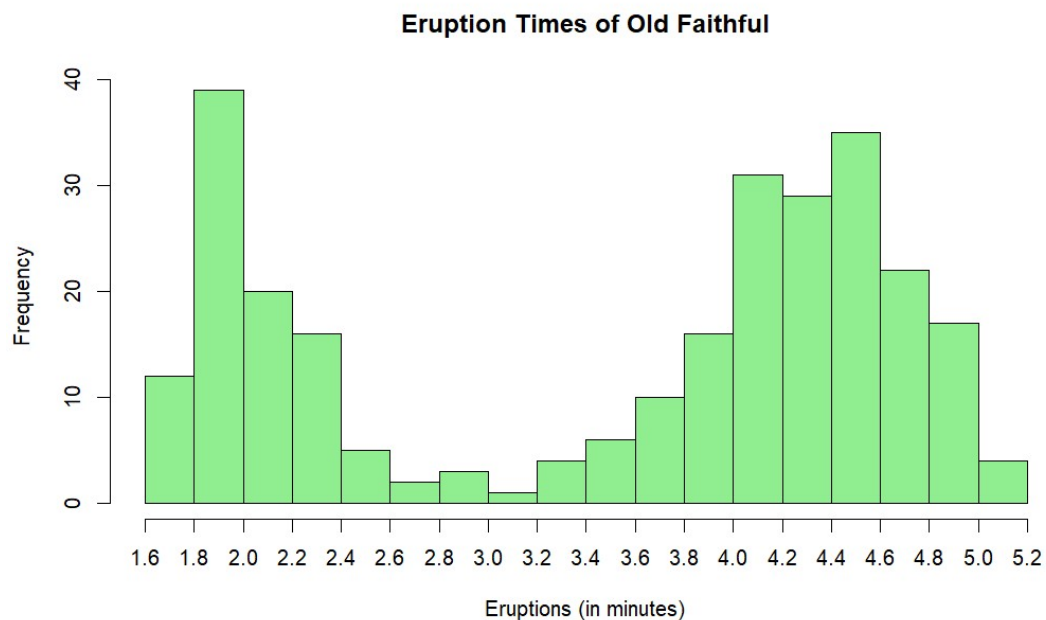
The first class contains all values from 1.600 to 1.799

The *class width* is:

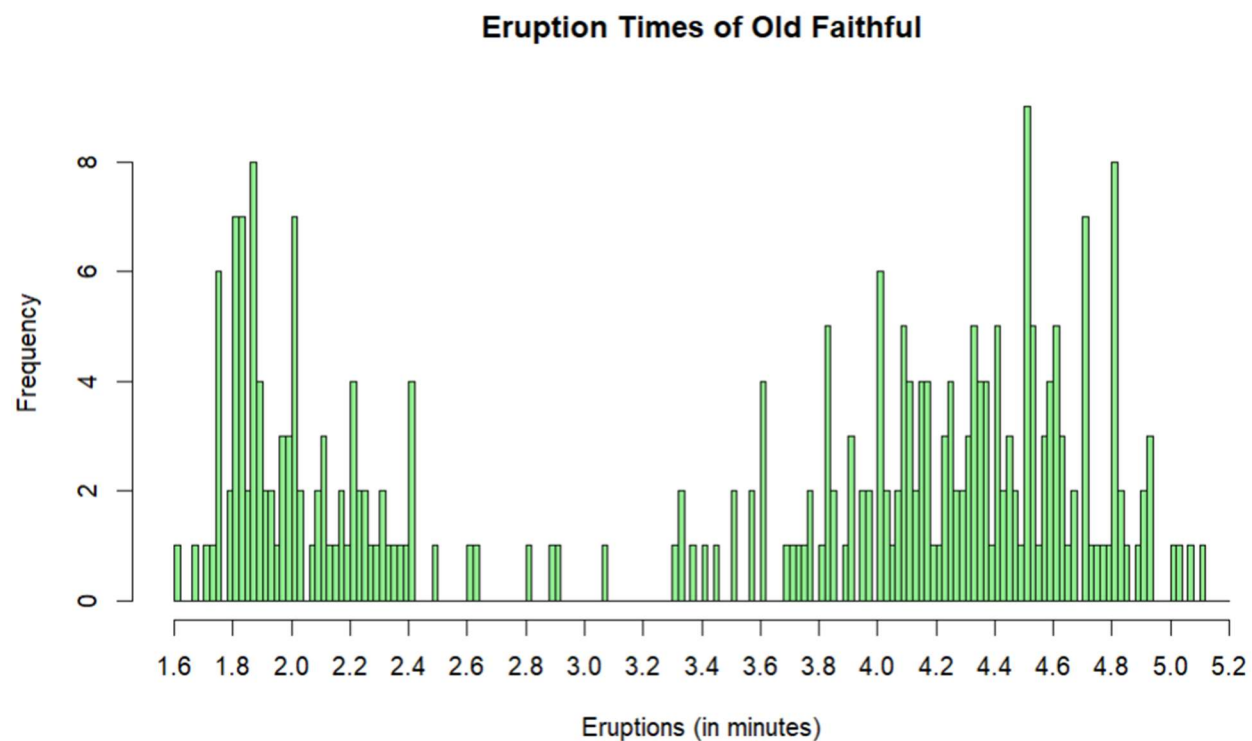
Lower Class Limit	Upper Class Limit	Frequency
1.600	1.799	12
1.800	1.999	39
2.000	2.199	20
2.200	2.399	16
2.400	2.599	5
2.600	2.799	2
2.800	2.999	3
3.000	3.199	1
3.200	3.399	4
3.400	3.599	6
3.600	3.799	10
3.800	3.999	16
4.000	4.199	31
4.200	4.399	29
4.400	4.599	35
4.600	4.799	22
4.800	4.999	17
5.000	5.199	4

A *histogram* is a direct visual representation of the frequency distribution.

```
> hist(faithful$eruptions)
```



Using too many classes is bad. For instance, here is the same data using a class width of 0.020 (instead of 0.200).



Cumulative Relative Frequency Distributions and Ogives

The *cumulative frequency* of a class is the number of individuals that fall into any class up to and including that class.

The *cumulative relative frequency* distribution is the *fraction* of individuals that fall into any class up to and including that class.

Lower Class Limit	Upper Class Limit	Frequency	Cumul Freq	Cumul Rel Freq
1.6	1.799	12		
1.8	1.999	39		
2.0	2.199	20		
... and so on ...				
5.0	5.199	4	272	

(complete table shown on next page)

Lower Class Limit	Upper Class Limit	Frequency	Cumul Freq	Cumul Rel Freq
	1.599		0	0.000
1.6	1.799	12	12	0.044
1.8	1.999	39	51	0.188
2.0	2.199	20	71	0.261
2.2	2.399	18	89	0.327
2.4	2.599	3	92	0.338
2.6	2.799	3	95	0.349
2.8	2.999	2	97	0.357
3.0	3.199	1	98	0.360
3.2	3.399	4	102	0.375
3.4	3.599	6	108	0.397
3.6	3.799	10	118	0.434
3.8	3.999	16	134	0.493
4.0	4.199	31	165	0.607
4.2	4.399	29	194	0.713
4.4	4.599	35	229	0.842
4.6	4.799	28	257	0.945
4.8	4.999	11	268	0.985
5.0	5.199	4	272	1.000
		Total = 272		

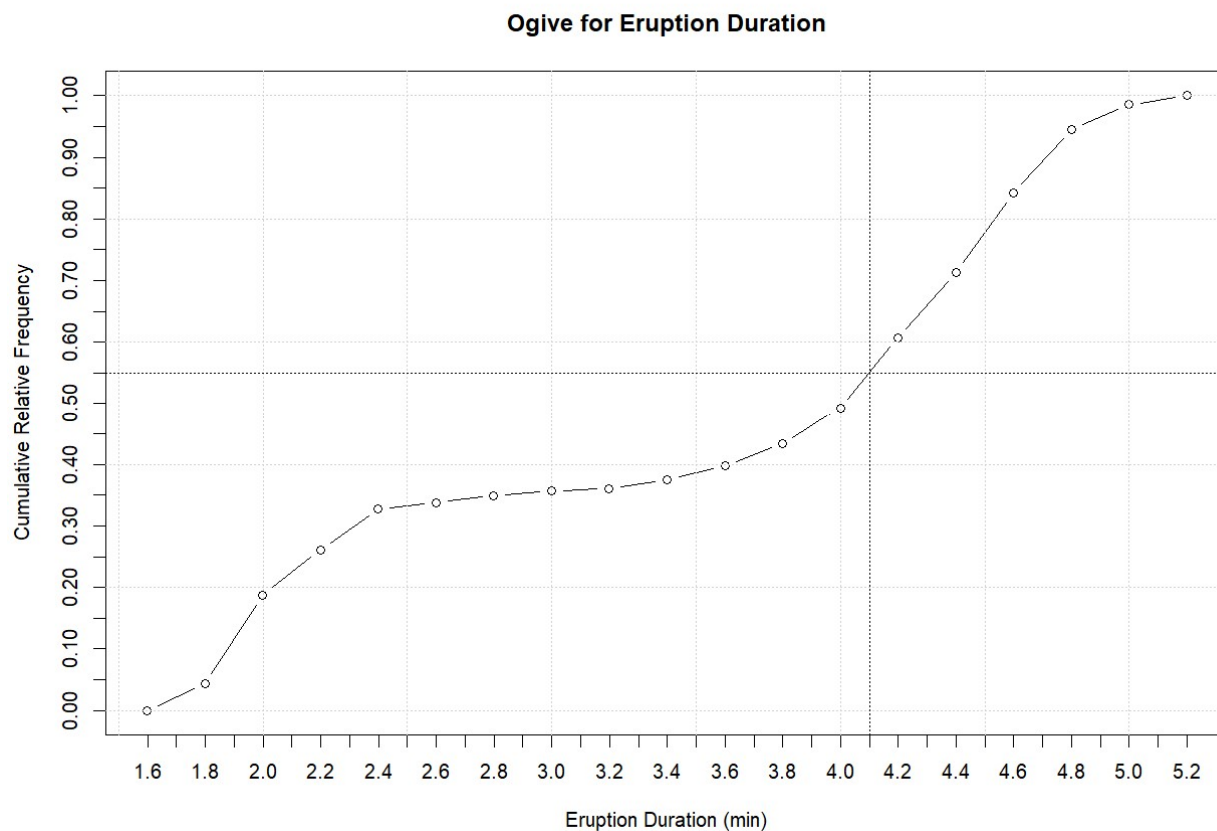
How many eruptions had a duration of less than 4.0 minutes?

What fraction of eruptions had a duration of less than 4.2 minutes?

What fraction of eruptions had a duration of less than 4.1 minutes?

An *ogive* (“oh-jive”) shows the *cumulative relative frequencies* plotted against the *upper class limits*.

(We insert one additional point at the beginning of the curve to indicate the 0.0 point of cumulative relative frequency.)



What percentage of eruptions have a duration below 3.0 minutes?

What percentage of eruptions have a duration below 4.1 minutes?

What value of eruption duration has 30% of eruptions below and 70% above?

Scatter Plot

When we measure two or more *numerical* variables for each individual/unit, then it is possible to create a *scatter plot*.

Supposing the two variables are called X and Y , then we simply plot one point for each individual/unit using the specific X and Y values as the coordinates.

A scatter plot helps us to quickly identify whether the variables X and Y are related.

Example (Fuel Efficiency vs Car Weight) Another built-in dataset called *mtcars* provides data for several variables for different makes of cars. If we let $Y = mpg$ (meaning fuel efficiency) and $X = wt$ (meaning weight), then the scatter plot shows the relation between fuel efficiency and car weight.

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear
Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4
Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4
Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4
Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3
Hornet Sportabout	18.7	8	360.0	175	3.15	3.440	17.02	0	0	3

Fuel Efficiency versus Weight for $n = 32$ cars

