

2 Organizing Data

2.1 Bar Graphs, Circle Graphs, and Time Plots

THE BIG IDEA: A picture is worth a thousand words. Once you've collected some data, it's time to describe it—and the best way to start describing data is with a picture.

BAR GRAPH: A graph that shows the possible values of a qualitative variable, and how often each value occurs. List the possible values of the variable along the x -axis (horizontally; left-to-right). Count how many times each value occurs, and make a bar for each value (as high as the number of times that it occurs).

EXAMPLE:

[1.] A report on road rage measured the day of the week that each of 69 incidents occurred. The results are shown below.

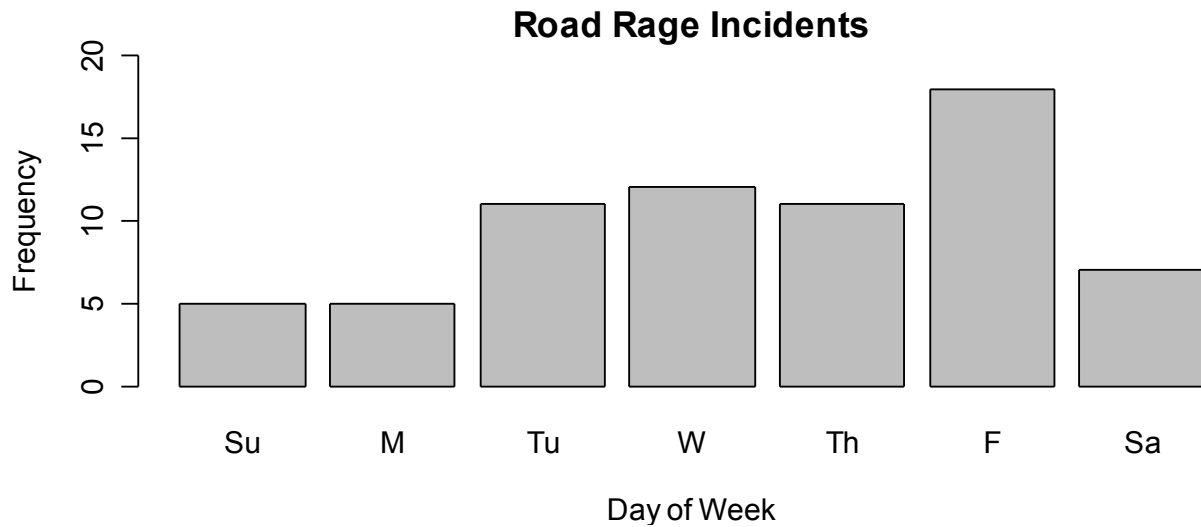
F	F	F	W	Sa	Tu	F
Tu	M	W	Su	Th	Th	W
Th	M	Th	Sa	F	Tu	Su
Tu	F	Th	F	Tu	F	W
F	F	Th	W	W	Tu	Tu
F	Th	Th	W	Su	Th	Sa
F	Tu	F	F	Su	F	M
F	Sa	Tu	W	Tu	F	Th
Sa	M	Th	W	Su	M	W
Sa	Tu	W	Sa	W	F	

Let's make a bar chart.

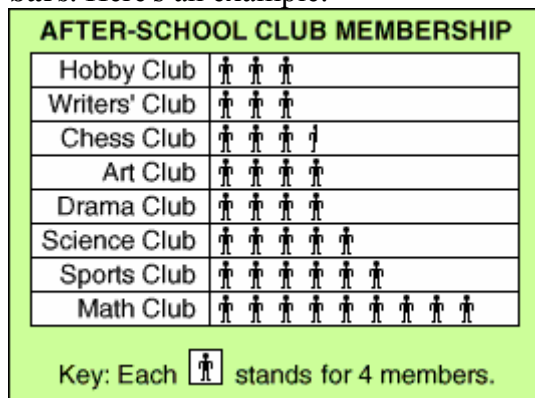
The possible values of the variable are Su, M, Tu, W, Th, F, Sa. Mark those along the horizontal axis. Now, let's count to see how many times each value occurs. The results are shown below.

Day	Su	M	Tu	W	Th	F	Sa
#	5	5	11	12	11	18	7

Friday occurs 18 times, so we need to mark the other axis (vertical) with numbers from 0 to 20. Finally, make a bar for each day. Here are the results:



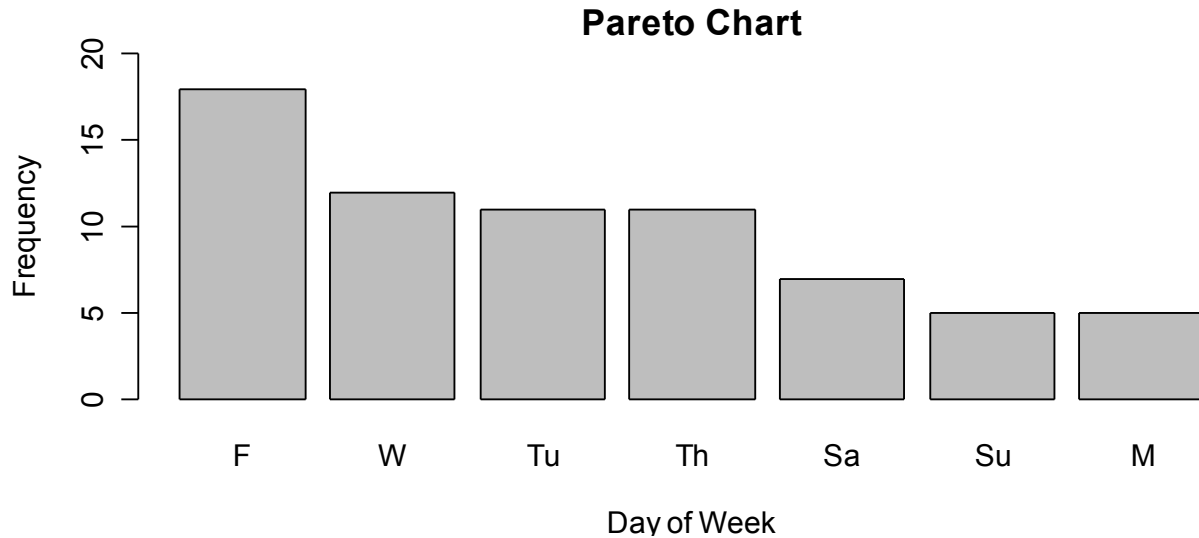
EXTRA! PICTOGRAPH: Like a bar chart, except that it uses stacks of pictures instead of bars. Here's an example.



PARETO CHART: A bar chart where the bars are ordered from tallest to shortest. Easy! This is often used to show the category that occurs most often (or least often).

EXAMPLE:

[2.] Using the road rage data...



PIE GRAPH: A graph that divides a circle into sectors (pie slices) based on the percentages for each value.

Just as with bar graphs, you should begin by counting how often each value occurs. The next step is to turn those counts into percents. Then, turn those percents into degrees. Finally, mark off angles for each value.

EXAMPLE:

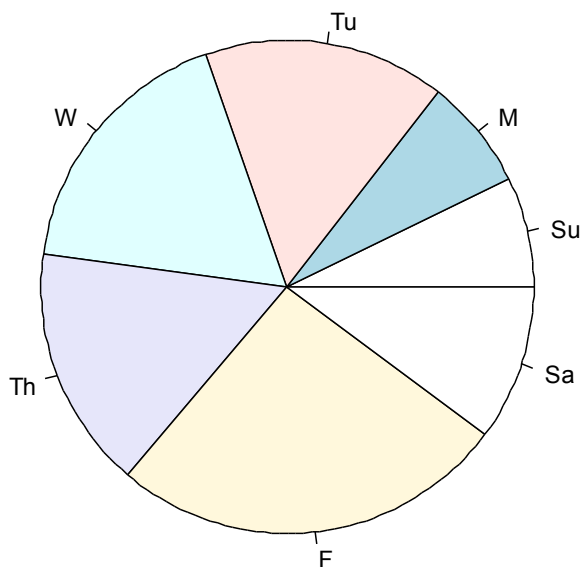
[3.] Continuing with our previous data—there were 69 observations. So, to turn each count into a percentage, we divide each count by 69. Actually that's not a percentage—it's a proportion—but it will do just fine. And less work is better, right?

Day	Su	M	Tu	W	Th	F	Sa
#	5	5	11	12	11	18	7
proportion	0.072	0.072	0.159	0.174	0.159	0.261	0.101

Now—since there are 360° in a circle, we need to take each of those numbers and multiply by 360° to get the angle size of each of the sectors.

Day	Su	M	Tu	W	Th	F	Sa
#	5	5	11	12	11	18	7
proportion	0.072	0.072	0.159	0.174	0.159	0.261	0.101
angle	26.1	26.1	57.4	62.6	57.4	93.9	36.5

So—starting from the positive x -axis, we make a 26.1° angle, which marks the boundaries of the "Sunday" pie slice. From there, mark another 26.1° angle for "Monday." Keep going...you should get something like this (without the fancy colors, of course).



There needs to be some way to show which sector goes with which value—in the chart above, notice the little line that leads to the abbreviation for the day.

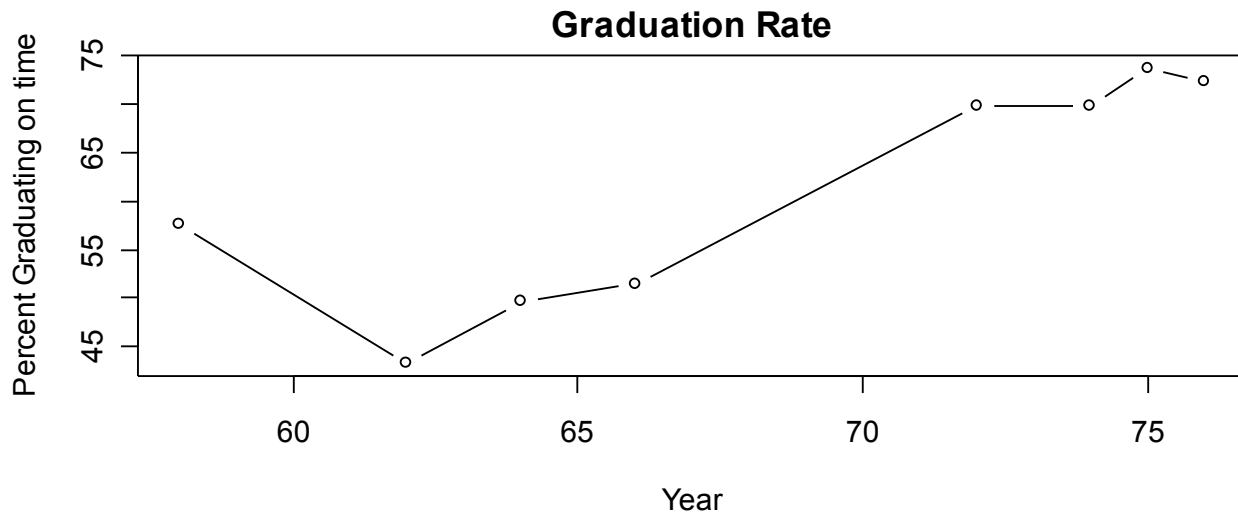
TIME PLOTS: Often, data are collected as time passes. If each measurement includes both a value and a time, then a time plot may be appropriate. **A time plot plots points, where the x -coordinate is the time, and the y -coordinate is the value.** Often, these dots are connected to make a sort of jagged line.

EXAMPLE:

[4.] The percentage of students who complete their coursework and receive a degree within 6 years is of some interest to universities. A major university tracked the percentage of freshmen who received their degree "on time" over the course of many years. The data are shown below.

Year	Percent
58	57.6
62	43.2
64	49.6
66	51.4
72	69.9
74	69.9
75	73.8
76	72.3

To make a time plot of this, first label the axes. The x -axis (horizontal) will be labeled with years—based on the data, we need years to go from about 1955 to 1980. The y -axis will be labeled for the other data—in this case, percentages from about 40 to 75. Plot and connect to get something like this:



2.2 Frequency Distributions and Histograms

HISTOGRAM: Where the bar chart is used for qualitative variables, the histogram is used for quantitative variables. **A histogram shows the possible values of the variable (divided into subranges), and how many data there are in each subrange.**

To construct a histogram, you must first create subranges in the data—groups; bins; buckets; there are lots of names (this is a fairly challenging step). Next, count how many data there are in each group. Finally, make a bar for each group.

Ideally, you want between 5 and 15 groups, so that your histogram will have between 5 and 15 bars. To do this, take the largest datum and subtract the smallest datum. Now divide that by 10. This number is how wide each group would need to be in order to have 10 groups—alas, it's probably a decimal—you don't want that. Round this (up or down) to some nice value that would be easy to count in multiples of. For example, it's easy to count by 5's, or 10's, but not easy to count by 7's. The number you get will be called the *group width*.

Now, look at the smallest datum you have. Round it *down* to the nearest multiple of the group width. This will be the *starting value*.

Let's continue with an example.

EXAMPLE:

[5.] Here are the final grades for some students in a statistics class.

88	82	89	70	85
63	100	86	67	39
90	96	76	34	81
64	75	84	89	96

Let's make a histogram.

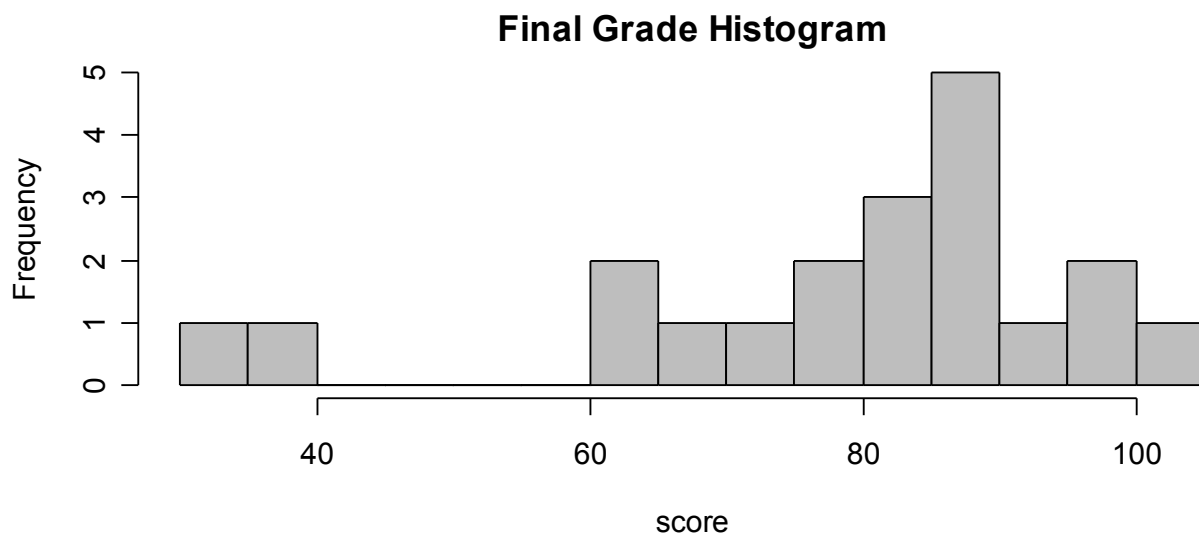
First of all, let's decide on a group width. The highest value is 100, and the lowest is 34. That's a difference of 66. Divide that by 10 and you get 6.6. Now, let's round that to a nice number—how about five? That's the closest nice number. *Our group width will be five.*

Now, the lowest number is 34—let's round that down to a multiple of five. The next lowest multiple of five is 30—*our starting value will be 30.*

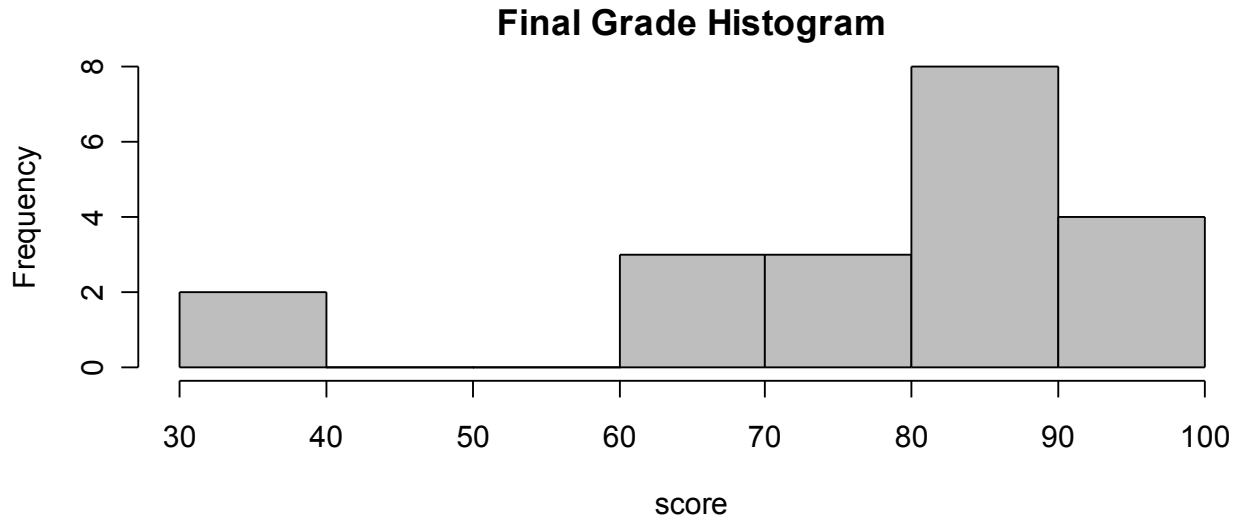
That means that the first group starts at 30; the second group starts at 35; the third group starts at 40; etc. Count how many data fall in each group. Here are the results:

At Least	Less Than	Count
30	35	1
35	40	1
40	45	0
45	50	0
50	55	0
55	60	0
60	65	2
65	70	1
70	75	1
75	80	2
80	85	3
85	90	5
90	95	1
95	100	2
100	105	1

And here is the histogram.



If you had rounded the group width to 10, here's the histogram you would get.



Both work well.

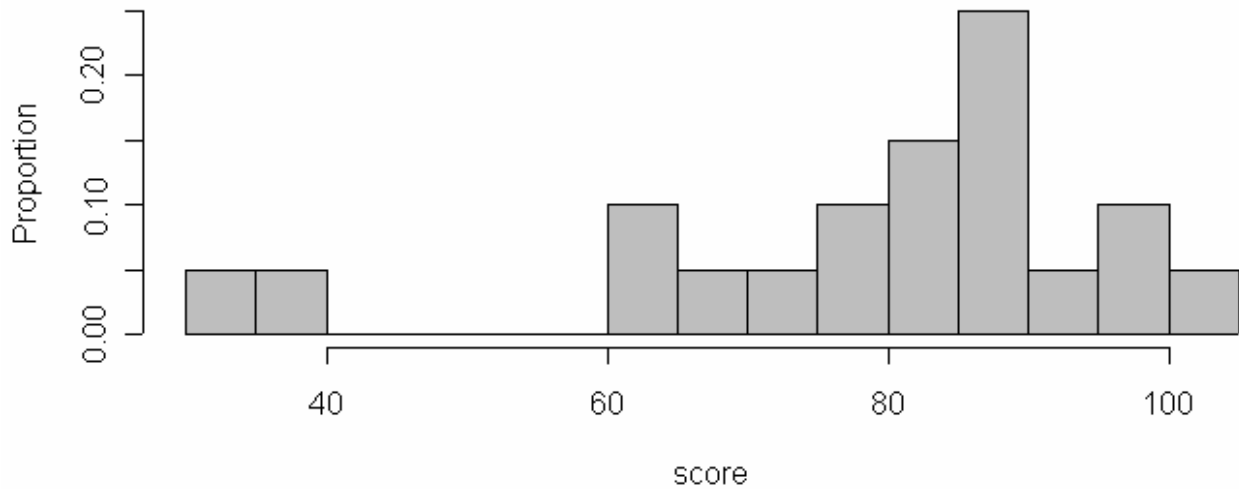
RELATIVE FREQUENCY HISTOGRAM: A histogram that shows the percentage of the data that are in each group. The process is the same, except that instead of just counting how many data are in each group, you find the percentage (or proportion) in each group by dividing by the number of data.

EXAMPLE:

[6.] Continuing our grade example—let's calculate the percentages.

At Least	Less Than	Count	%
30	35	1	5
35	40	1	5
40	45	0	0
45	50	0	0
50	55	0	0
55	60	0	0
60	65	2	10
65	70	1	5
70	75	1	5
75	80	2	10
80	85	3	15
85	90	5	25
90	95	1	5
95	100	2	10
100	105	1	5

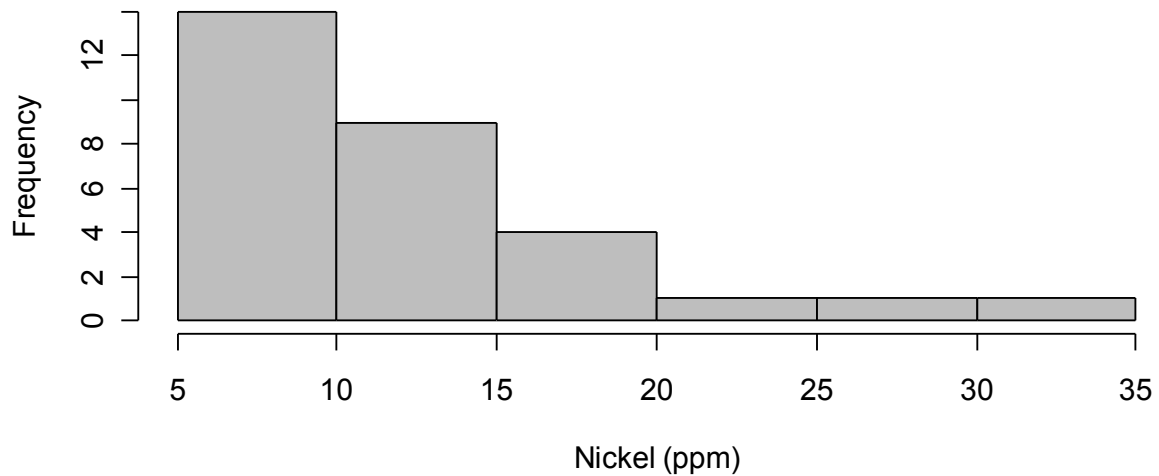
Final Grade Histogram



THE SHAPE OF THE DATA: One of the main features that we look at in a histogram is shape. There are lots of named shapes out there, but there are really only two that are of sufficient importance for us—symmetric, and skewed. A **Symmetric** shape is just what it sounds like—symmetric!

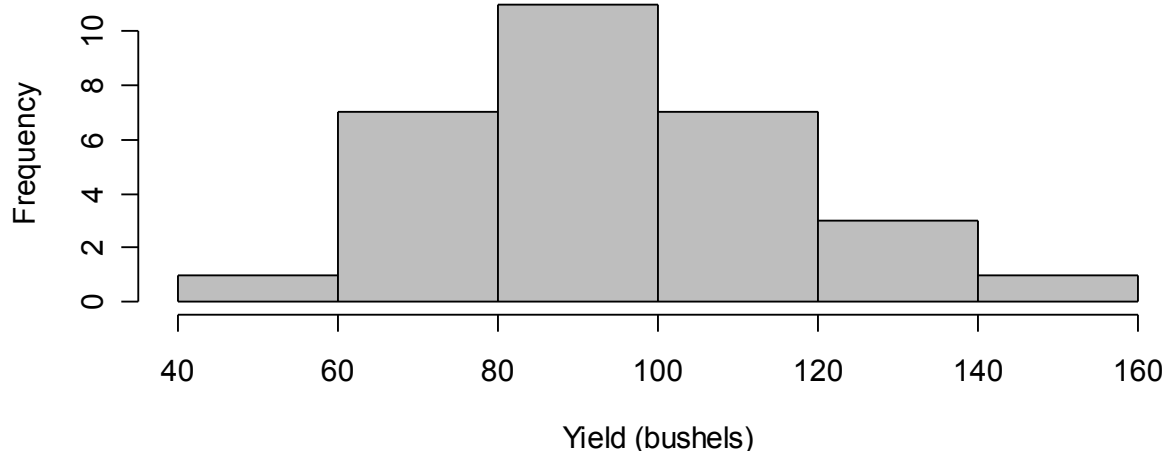
If the shape isn't symmetric, and it's higher on one side than the other, then it's **skewed**. In particular, the side that's lower is the side that's skewed. Here are some sample pictures:

Nickel Concentration



The histogram above is skew right.

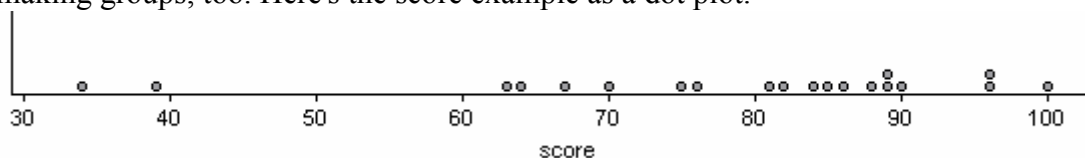
Barley Yield, 1932



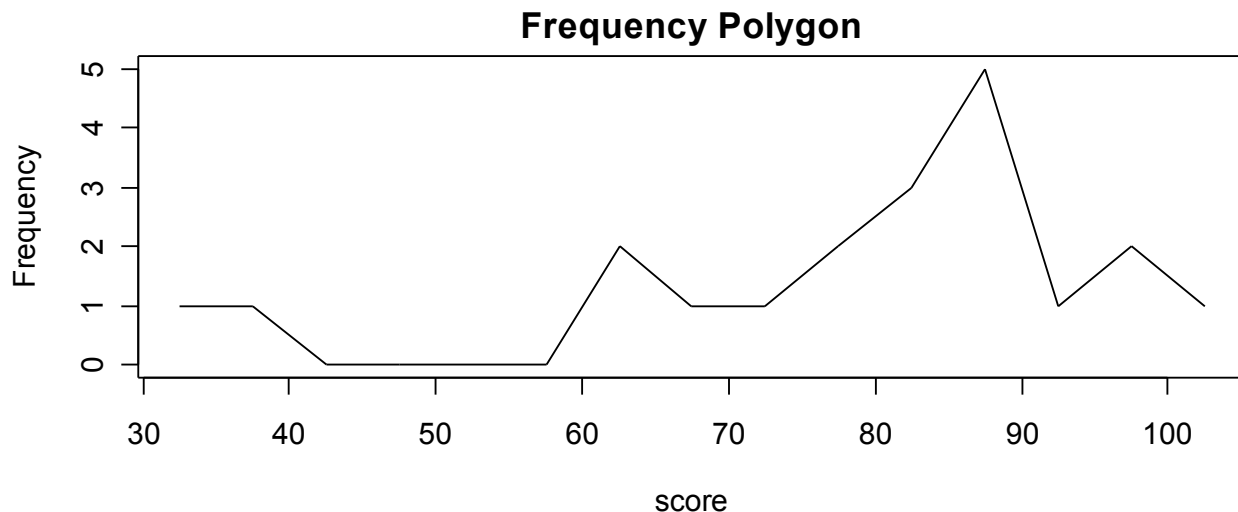
The histogram above is symmetric.

EXTRA!

DOTPLOT: A histogram using stacks of dots rather than bars. There's less emphasis on making groups, too. Here's the score example as a dot plot.



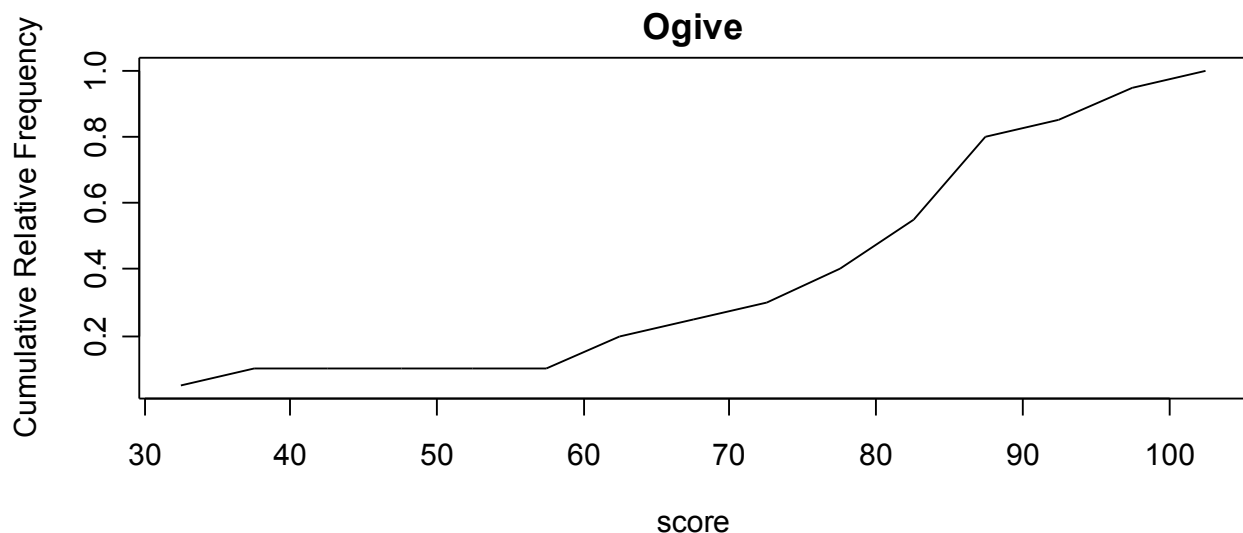
FREQUENCY POLYGON: This shows the same thing as a histogram, but uses a connected line rather than separate bars. Think of it as a cross between a dot plot and a time plot. Create groups just as if you were going to make a histogram, but just plot a single dot at the height for each group—horizontally, the dot is placed at the midpoint of the group. Here's the score example (once again) as a frequency polygon.



CUMULATIVE RELATIVE FREQUENCY POLYGON (OGIVE): A Frequency Polygon that shows the cumulative relative frequency for each group. The cumulative frequency for a group is how many data are in that group AND all previous groups. The cumulative relative frequency for a group is the relative frequency (percent) for that group AND all previous groups. Here's the chart for our score example. Note that the last entry is 100—this is always the case!

At Least	Less Than	Count	%	cumulative %
30	35	1	5	5
35	40	1	5	10
40	45	0	0	10
45	50	0	0	10
50	55	0	0	10
55	60	0	0	10
60	65	2	10	20
65	70	1	5	25
70	75	1	5	30
75	80	2	10	40
80	85	3	15	55
85	90	5	25	80
90	95	1	5	85
95	100	2	10	95
100	105	1	5	100

And here's the graph.



2.3 Stem-and-Leaf Displays

STEMPLOT: The Stemplot (or Stem-and-Leaf) is like a histogram turned on its side, with numbers instead of bars. Let's start with a stem-and-leaf of the score data that we've been using—and I'll drop in the histogram for comparison.

```

3 | 49
4 |
5 |

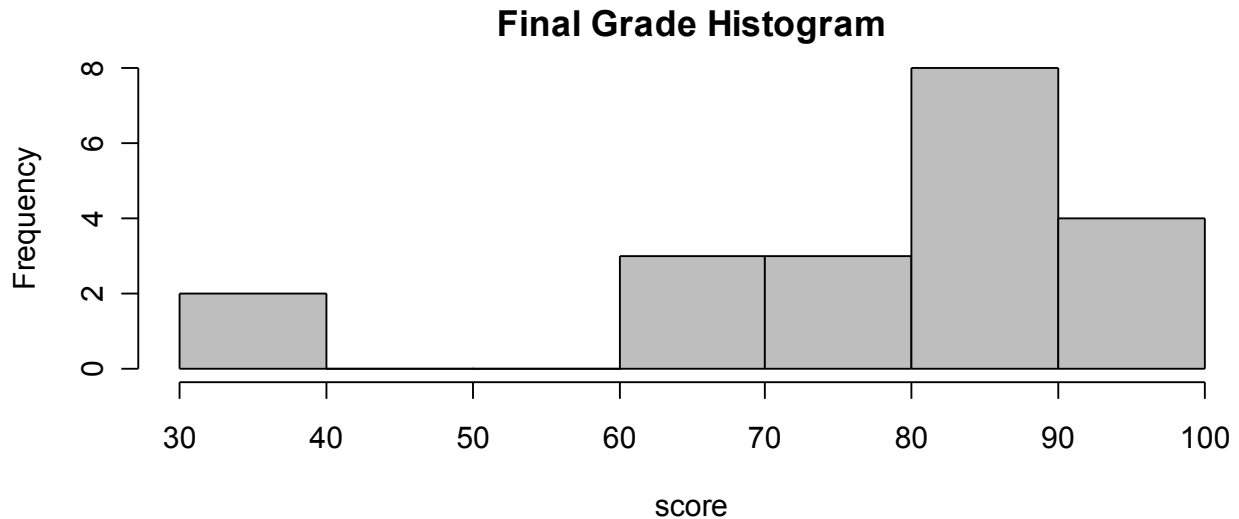
```

```

6 | 347
7 | 056
8 | 12456899
9 | 066
10| 0

```

where 10|0 means a score of 100



Turn your head to the right when you look at the stemplot—it looks a little like the histogram, doesn't it? That's the beauty of the stemplot!

OK, here's the deal—with a stemplot, the group width must be some power of ten. In other words, the groups are based on *place value*. The groups are on the left, and are called **stems** in this display. The numbers on the right are called **leaves**. A stem and a leaf join to form one of the numbers in the data set.

Notice that all of the leaves are single digits. Write all of the data using the same number of digits. The last (right-most) digit becomes the leaf, so all of the other digits form the stem. Look at your data; decide what stems are needed, and write them down; then collect the leaves. That's it!

EXAMPLE:

[7.] Short-term investments are methods (groups, funds, etc.) that take a sum of money and return that money with interest in a relatively short time (as little as one month). The following are the maturity times (how long you've got to wait to get your money back) in days for 40 short-term investments.

70	62	99	85	51	55	57	75	60	64
64	38	68	79	36	81	53	56	69	89
99	67	95	83	63	80	47	71	78	87
55	70	86	70	66	98	50	51	39	65

Let's make a stemplot.

First of all, do the data all have the same number of digits? Yes; good.

Since all of them have 2 digits, the stems will be single digits, like 7, or 6, or 9—the tens digit of each number is its stem.

Look at the data. What kinds of stems will be needed? In particular, what are the lowest and highest stems needed?

The smallest number is 36, which has a stem of 3. The biggest number is 99, which has a stem of 9. So—the first thing we should do is create a line of stems, from 3 to 9.

```
3 |  
4 |  
5 |  
6 |  
7 |  
8 |  
9 |
```

Now, start going through the data and listing the leaves. Here's what I got after the first ten data.

```
3 |  
4 |  
5 | 157  
6 | 204  
7 | 05  
8 | 5  
9 | 9
```

Keep going!

```
3 | 869  
4 | 7  
5 | 15736501  
6 | 2044897365  
7 | 0591800  
8 | 5193076  
9 | 9958
```

Typically, the leaves are then sorted from smallest to largest along each stem—which gives you this:

```
3 | 689  
4 | 7  
5 | 01135567  
6 | 0234456789  
7 | 0001589  
8 | 0135679  
9 | 5899
```

That leaves (not a pun—really!) only one thing—the legend. This plot is based on place value, but without a legend, the reader cannot know what the place values are! 3 | 6 might mean 36, or 3.6, or 0.36!

```
3 | 689  
4 | 7  
5 | 01135567
```

6|0234456789
7|0001589
8|0135679
9|5899

where 3|6 means 36 days

EXTRA! That leaves just two problems—first, what happens when there are too few stems?

EXAMPLE:

[8.] The cholesterol levels of 20 young patients were taken by a pediatrician. Here are the data:

210	202	215	208	217
209	218	221	210	207
212	200	213	210	210
208	214	218	199	203

The data all have three digits—the stems will be the first two digits of each number.

But that only gives us 19, 20 and 21!

To fix this, we'll list every stem twice, but only put half of the leaves on each stem. **This is called splitting the stems.**

19|
19|
20|
20|
21|
21|
22|
22|

In particular, leaves of zero through four will go on the first stem, and five through nine will go on the second stem. Here's the plot with the first five data:

19|
19|
20|2
20|8
21|0
21|57
22|
22|

OK, here's the finished plot.

19|
19|9
20|023
20|7889
21|0000234
21|5788
22|1

where 19|9 means a cholesterol level of 199

In this example, we split each stem into two parts—note that it is also possible to split stems into five parts.

As for the second problem—what if there are too many stems?

EXAMPLE:

[9.] A sample of 13 batches of Portland cement were measured for the heat emitted (calories per gram). Here are the data:

78.5	74.3	104.3	87.6	95.9	109.2	102.7
72.5	93.1	115.9	83.8	113.3	109.4	

The stems are all digits except the rightmost—for these data, that's from 72 to 113! That's way too many! To fix this, we must **round** each datum one place value—to the nearest unit (one) in this case.

79	74	104	88	96	109	103
73	93	116	84	113	109	

Now the stems are 7 through 11—just enough!

7 | 349

8 | 48

9 | 36

10 | 3499

11 | 36

where 11|6 means 116 calories per gram

EXTRA!

BACK-TO-BACK STEMLOTS: Often, we want to compare two sets of data. You can make two histograms, or two dotplots—two of anything! The hitch is that you should use the same scales on both in order to make comparisons easy. However, because of the way they're made, two stemplots are a better choice. The reason is that you can share the stems between the two plots, which ensures a common scale. **This technique is called back-to-back stemplots.**

EXAMPLE:

[10.] One of the statistics that the government keeps is the labor force participation rate, a measure of how many women are working. Here are the numbers for a selection of cities during the year 1968:

0.50	0.54	0.43	0.42
0.58	0.42	0.55	0.50
0.49	0.51	0.45	0.52
0.56	0.49	0.34	0.45
0.63	0.54	0.45	

And here are the numbers for the same cities in 1972:

0.59	0.55	0.46	0.45
0.64	0.52	0.55	0.50
0.50	0.53	0.60	0.52
0.57	0.57	0.49	0.45

0.64 0.53 0.35

Let's make a back-to-back stemplot to compare the two.

All of the data have the same number of digits; the stems are 3, 4, 5 and 6—that's not quite enough, so let's split the stems. The stems will go down the middle of the plot, and leaves will point out in both directions from the stems (one direction for the 1968 data; the other direction for the 1972 data).

1968		1972
4	3	
	3	5
322	4	
99555	4	5569
442100	5	0022333
865	5	55779
3	6	044
	6	

where 3|6|0 means 0.63 in 1968 and 0.60 in 1972.

Notice that the leaves on the left go backwards—when you put the leaves in order, you do so by pointing *away* from the stems.