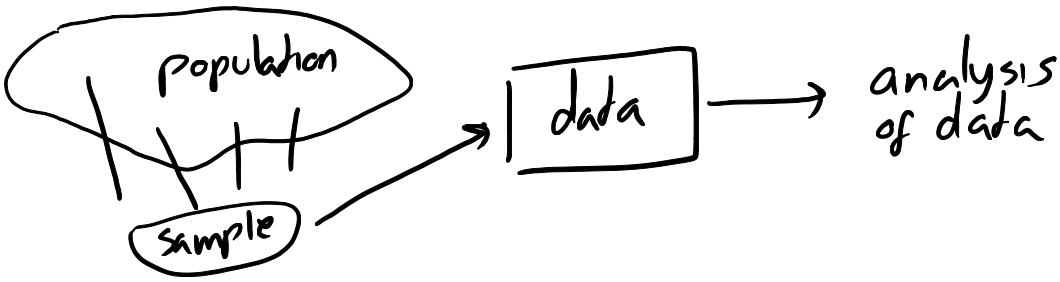
# Making data plots (to learn something about data)




## Histogram

A graph consists of bars of equal width drawn adjacent to each other (unless there are gaps in the data). e.g. world record 100m dives for men and women

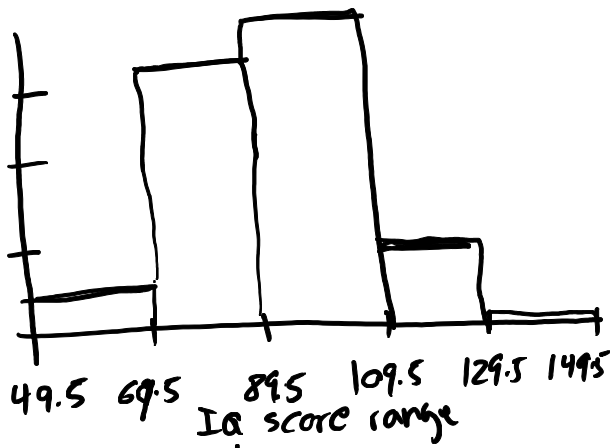
horizontal scale: classes of quantitative data values

vertical scale: represents frequencies

class width $= 69.5 - 49.5 = 20$

Ex)

| IQ Score | Frequency |
|---|---|
| 50-69 | 2 |
| 70-89 | 33 |
| 90-109 | 35 |
| 110-129 | 7 |
| 130-149 | 1 |




A histogram is a graph of a frequency distribution

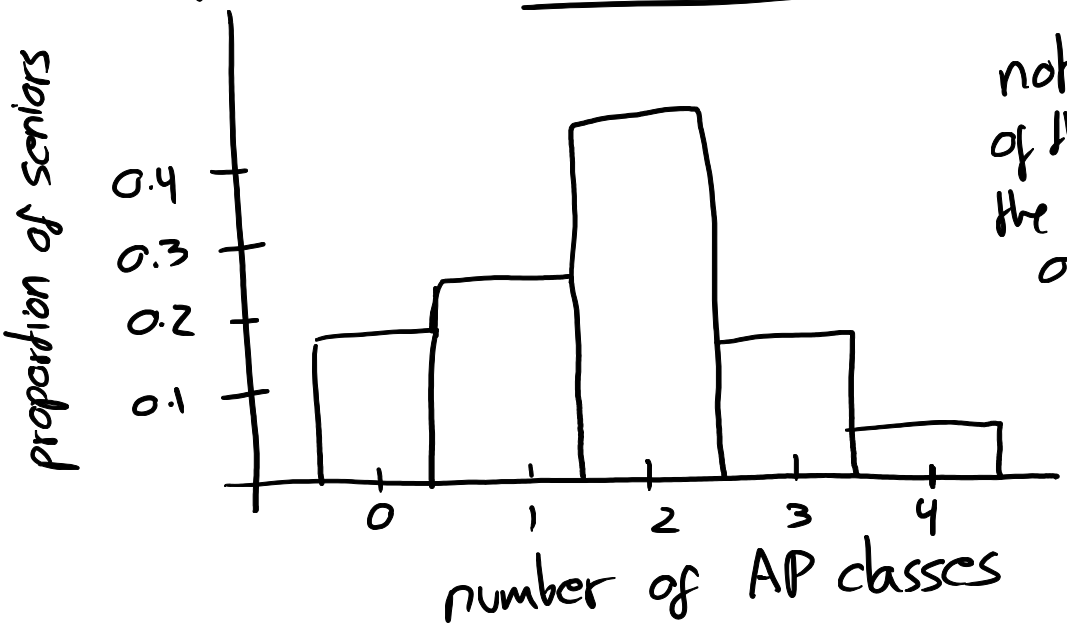
Relative frequency histogram: has the same shape and horizontal scale as a histogram but the vertical scale is marked with relative frequencies instead of actual frequencies.

Ex)

| relative cum freq. | # of AP classes taken by sr. year | Frequency | Relative frequency |
|---|---|---|---|
| 0.18 | 0 | 400 | 400/2200 = 0.18 |
| 0.41 | 1 | 500 | 500/2200 = 0.23 |
| 0.82 | 2 | 900 | 900/2200 = 0.41 |
| | 3 | 300 | 300/2200 = 0.14 |
| ⋮ | 4 | 100 | 100/2200 = 0.05 |

adds to 2200/2200

recall from last class: cumulative frequency



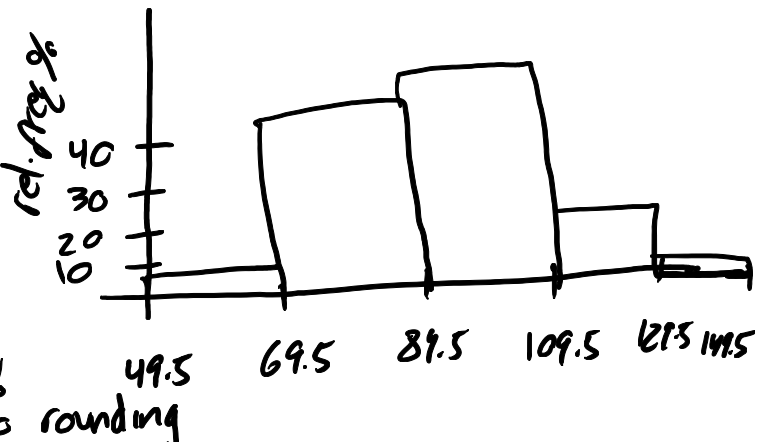

note that shape of this histogram is the same as the previous one.

note that sum of relative frequencies is 1.01 due to roundoff to 2 decimal digits.

| IQ score | relative freq |
|---|---|
| 50-69 | 2.6% |
| 70-89 | 42.3% |
| 90-109 | 44.9% |
| 110-129 | 9.0% |
| 130-149 | 1.3% |

adds to 100.1% due to rounding

class width = 69.5 - 49.5 = 20



objective: not to simply graph the histogram but to understand something about the data.

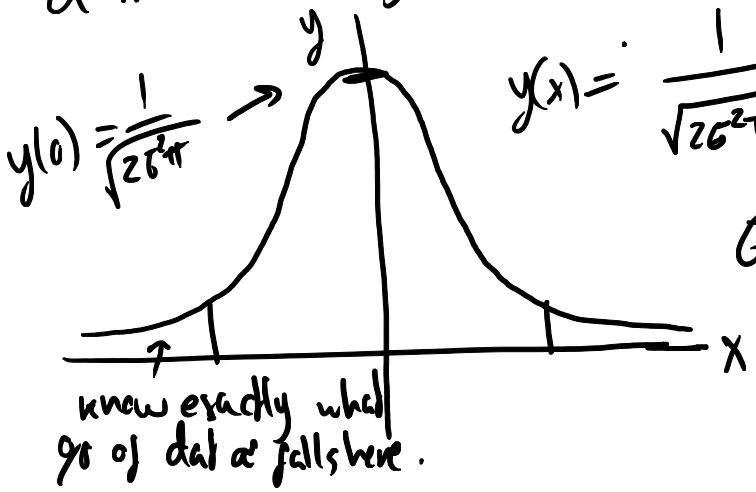
Q: does the histogram follow a "bell" shape?

A <u>normal</u> distribution has a "bell" shape.

Characteristics of bell shape: (Gaussian distribution)

(1) The frequencies increase to a maximum, and then decrease

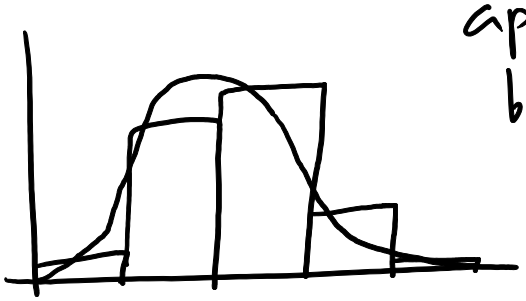
(2) symmetry, with left half of the graph roughly a mirror image of the right half.




$$y(x) = \frac{1}{\sqrt{2\sigma^2\pi}} \exp\left[-\frac{x^2}{2\sigma^2}\right]$$

$\sigma^2 > 0$ is the <u>variance</u>

Ex) IQ scores



bell curve approx. fits the histogram

approximately bell shaped distribution

why important? b/c then we can use results regarding the normal distribution.

later in course: we will study <u>central limit theorem</u>

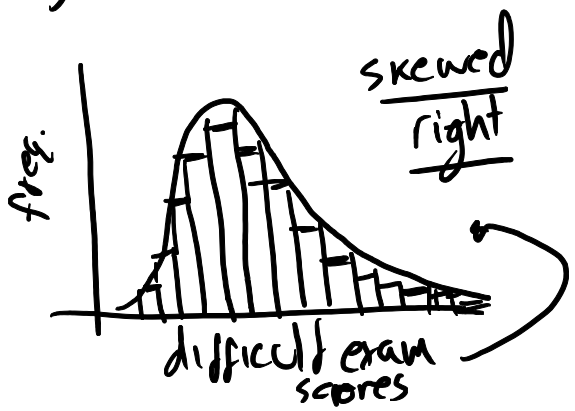
under some assumptions, the distribution of a large sample will always approach a Gaussian one.

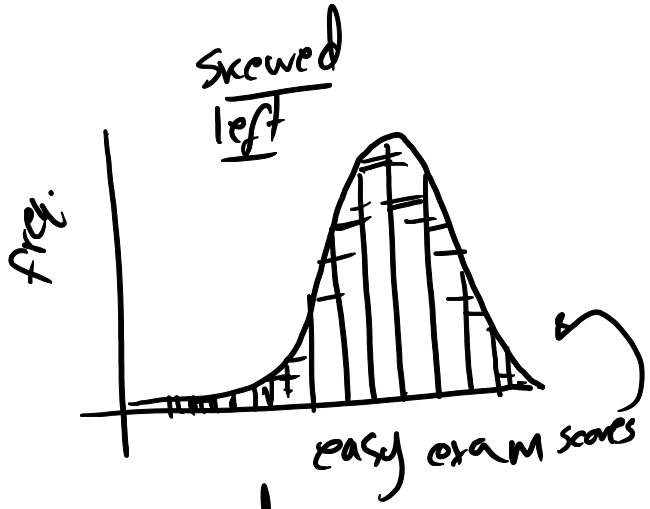
## Skewness

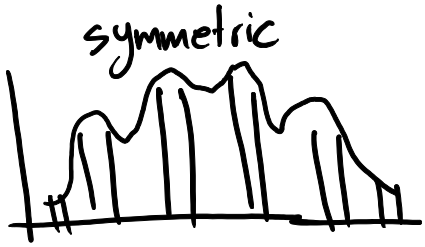
A distribution of data is skewed if it is not symmetric and extends more to one side than to the other.

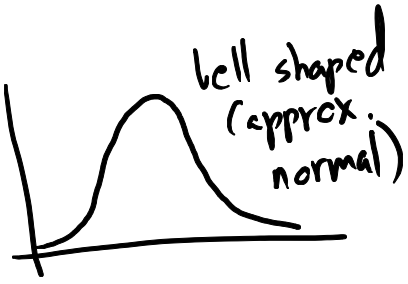
Data skewed to the right (positively skewed) have a longer right tail.

Data skewed to the left (negative skewed) have a longer left tail.



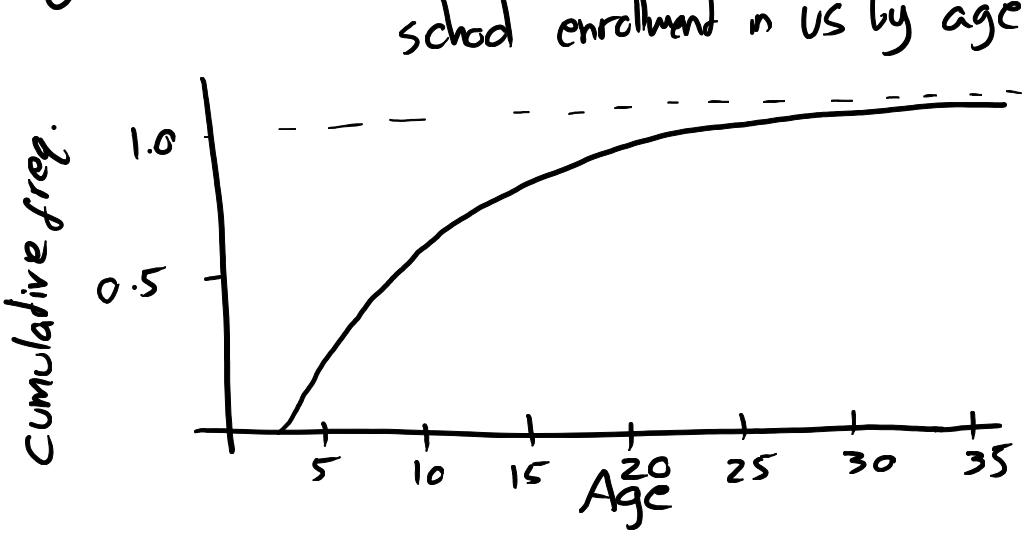





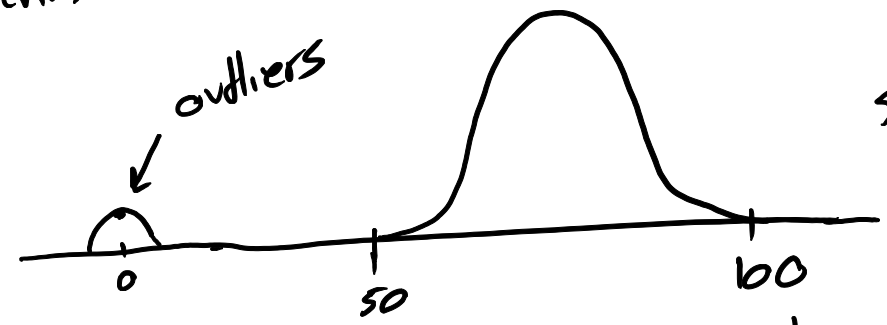
Sometimes we sum frequencies and show the result visually in a cumulative relative frequency plot.



what can we learn from such a plot. ex: 95% of enrollment is below age 30, and thus, 5% is over age 30.

A distribution skewed to the left has a cumulative frequency plot that rises slowly at first and becomes steeper later.

Outliers Extreme values, called outliers, are found in many distributions. Ex) distribution of exam scores of 100 students. 3 students missed the exam and got a zero.




outliers can skew statistics if one is not careful.

often result of errors in measurement or special situations: deserve scrutiny.
However, outliers can also be the result of natural chance variation. (ex) how much excited electrons jump )

Summarizing distributions (measuring center, spread) variation

median : middle number of a data set arranged in numerical order.

mean: average (much more susceptible to outliers!)

ex) {387, 400, 400, 410, 410, 410, 414, 415, 420, 420, 421, 457, 461}

median: 414

mean: $\frac{\Sigma \text{values}}{13} = 417.3$

} in this case the mean and median are very close.

but consider some exam scores:

80, 90, 95, 100, 0

missed the exam (outlier)

0, 80, (90), 95, 100

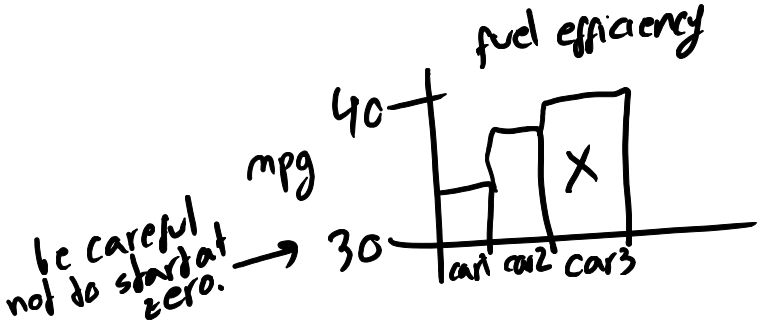
↓ median

but mean = $\frac{365}{5} = 73 < 80$

mean has been brought down by the zero outlier. not representative of those who took the test.

key points:

1. watch data for outliers
2. watch to see if data is skewed or approximately bell shaped.
3. In plots, make sure that the chosen y range makes sense.




car3 is not twice as good as car 1 as the table seems to indicate.

## Pulse rates of males (2.3.11) uses ex (2.2.21)

60, 74, 86, 54, 90, 80, 66, 68, 68, 56, 80, 62, 74, 60, 52, 60, 66, 64, 64, 46, 68, 58, 68, 70, 56, 66, 78, 68, 62, 70, 72, 74, 64, 50, 70, 58, 60, 88, 84, 76

Begin with lower class limit of 40 and use a class width of 10.

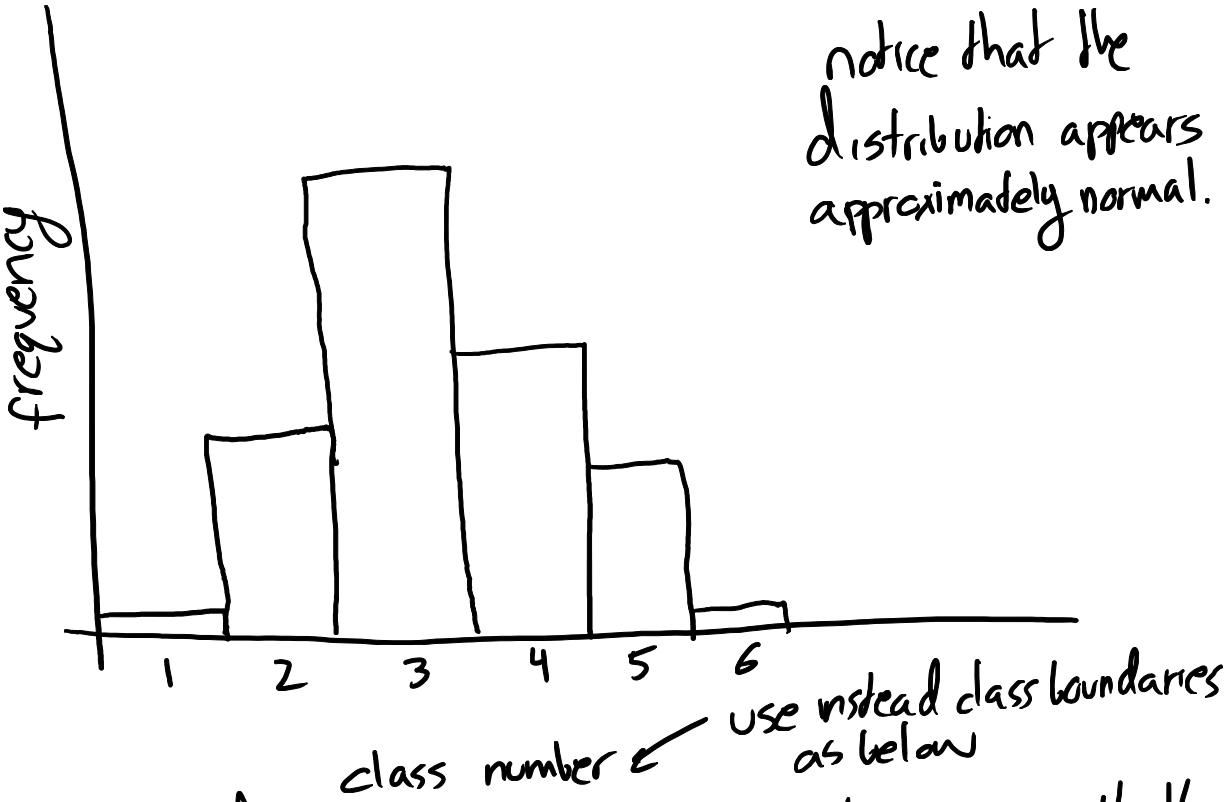
| | range of class |
|---|---|
| class 1: | 40-49 |
| class 2: | 50-59 |
| class 3: | 60-69 |
| class 4: | 70-79 |
| class 5: | 80-89 |
| class 6: | 90-99 |

} classes based on range of 46-90 class width 10 and lower class limit 40

Next, count how many of the numbers fall in each class.

| class # | frequency |
|---|---|
| 1 | 1 |
| 2 | 7 |
| 3 | 17 |
| 4 | 9 |
| 5 | 5 |
| 6 | 1 |

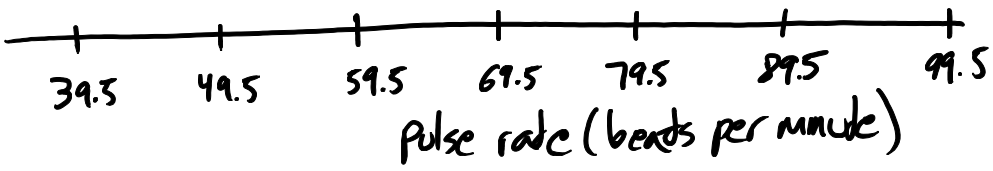
Based on the data, we make the following histogram:




notice that the distribution appears approximately normal.

use instead class boundaries as below

However, instead of class number we label the x-axis with the boundaries of the classes:




Let's look now at another example:

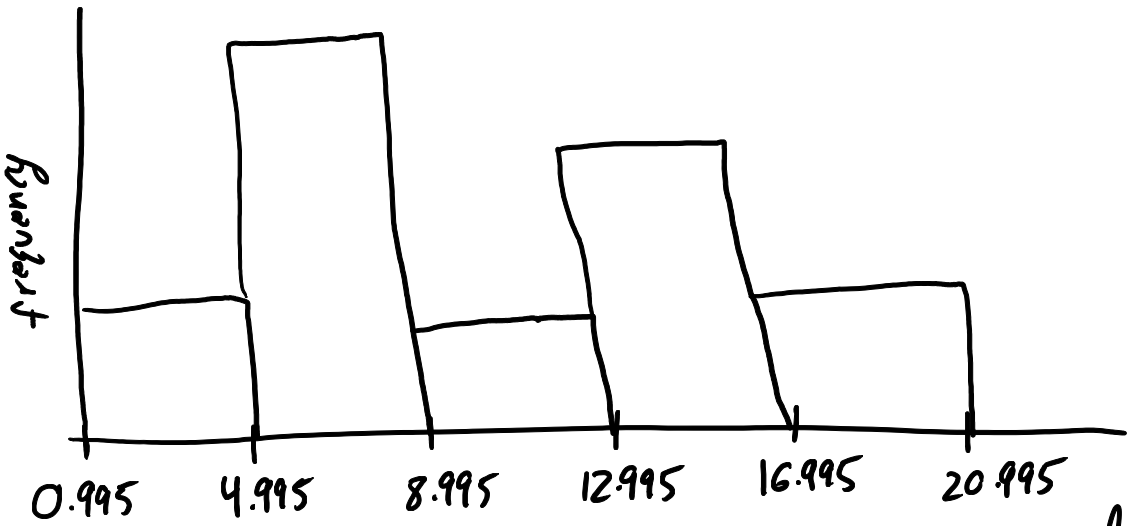
Ex) (2.3.14) uses data set specified in (2.2.24)
Data set 16, appendix B. Begin with a lower class limit of 1.00km and use a class width of 4.00km.

The following classes are chosen

| | |
|---|---|
| class 1 : | focus depth (km) 1.00 – 4.99 |
| class 2 : | 5.00 – 8.99 |
| class 3 : | 9.00 – 12.99 |
| class 4 : | 13.00 – 16.99 |
| class 5 : | 17.00 – 20.99 |

| depth (km) | frequency |
|---|---|
| 1.00 - 4.99 | 7 |
| 5.00 - 8.99 | 21 |
| 9.00 - 12.99 | 4 |
| 13.00 - 16.99 | 12 |
| 17.00 - 20.99 | 6 |

The following histogram then results:




This histogram is not bell shaped; distribution does not appear normal.