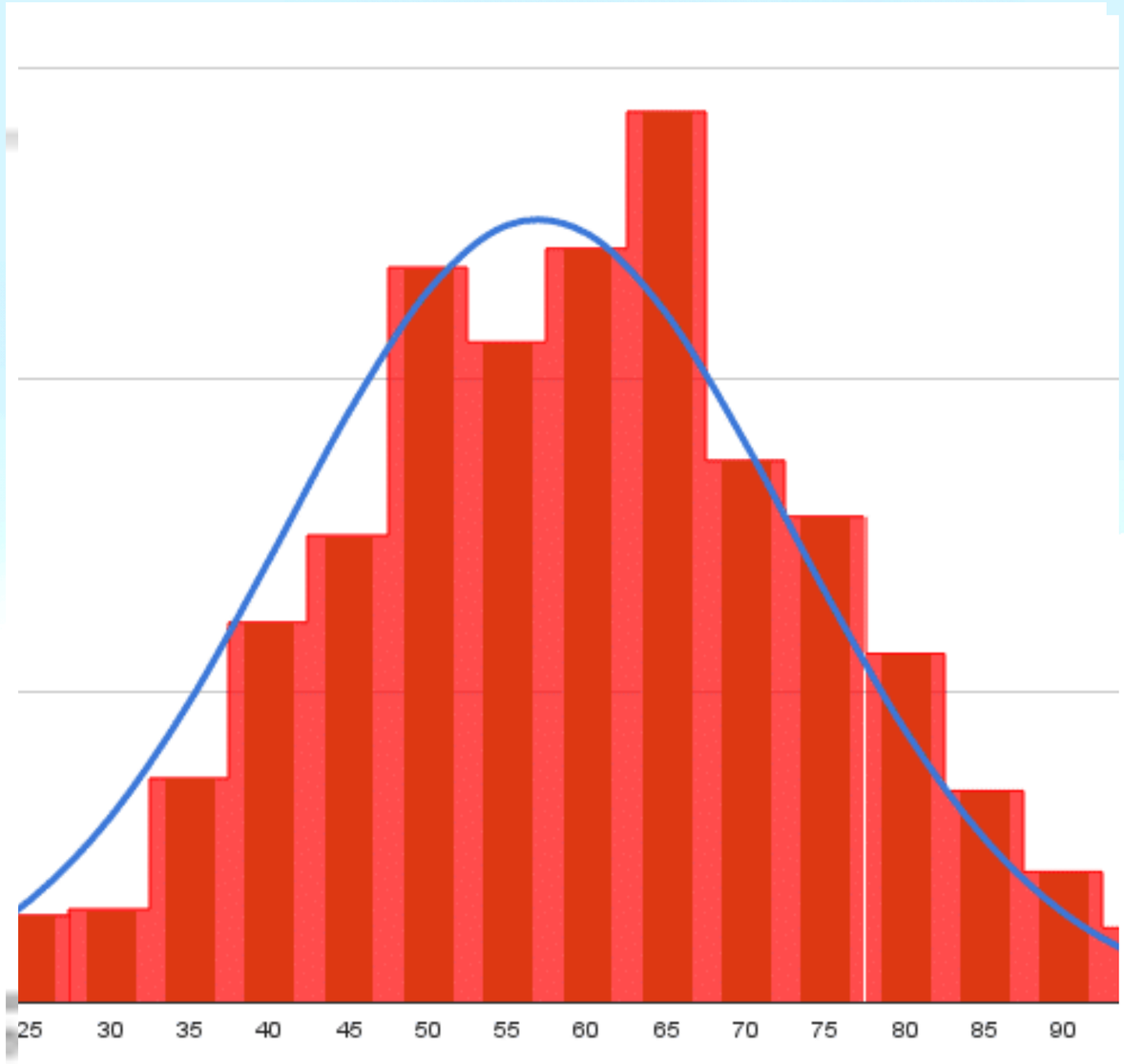
Distributions

PHYS 2601

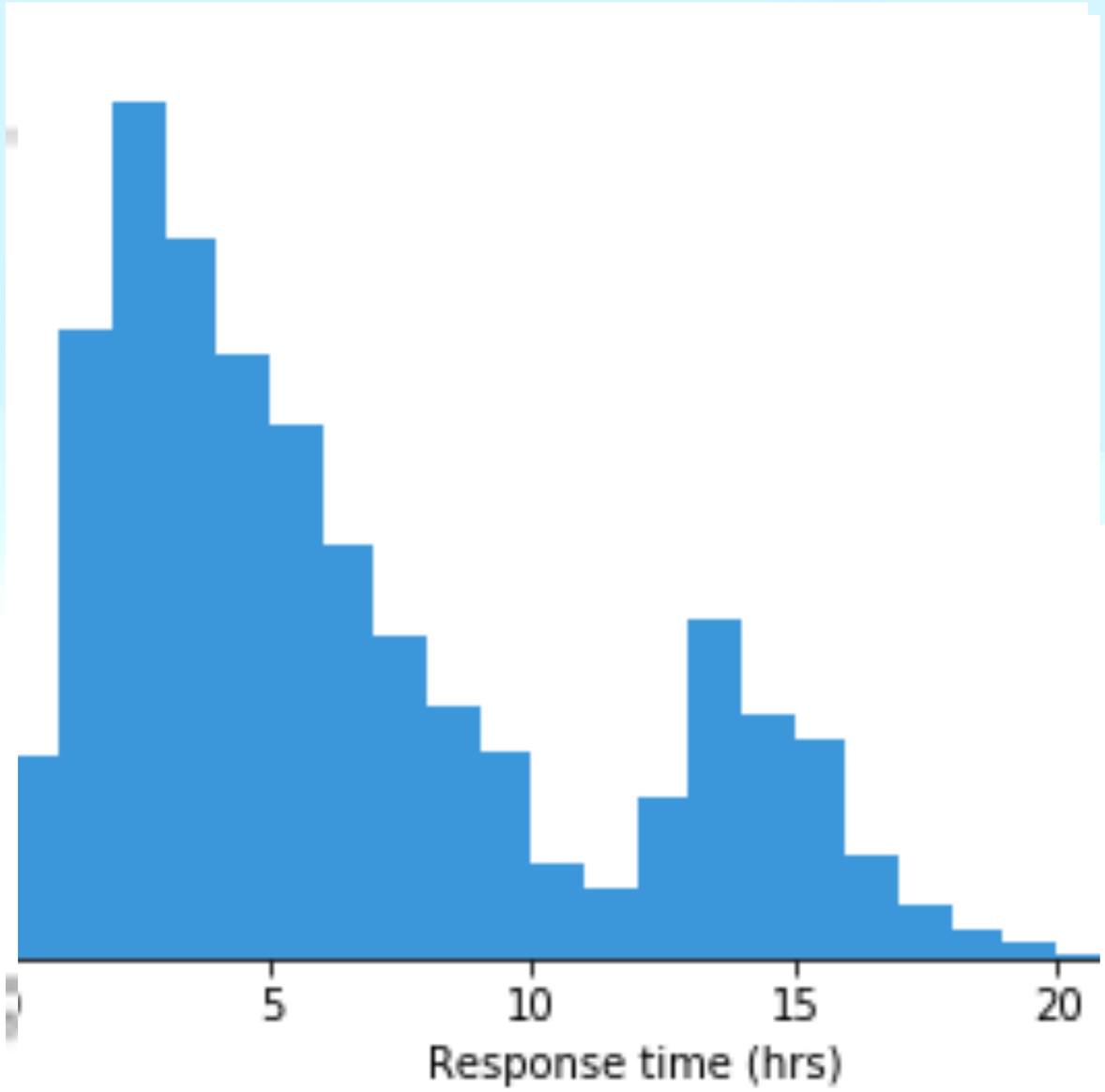
Multiple Measurements

- If possible, the best thing to do about measurement errors is to just take many many measurements of the same thing.
- This is not always possible, but when it is it allows one to calculate the precision of ones measurement.
- Note that knowing the precision, doesn't tell you the accuracy. So you know how close your measurements are to each other, but not how close they are to the ***true*** value.



Histograms

- A histogram plots the number of occurrences of a value versus the value.
- If one has discrete values, this can be the number of 5s, the number of 12s.
- If continuous, then the number of occurrences in a bin. The number of values > 5.8 and < 6.0. The histogram will change if you change the binning.



Histograms

- The y-axis of the plot can be number of occurrences per bin

$$n_i$$

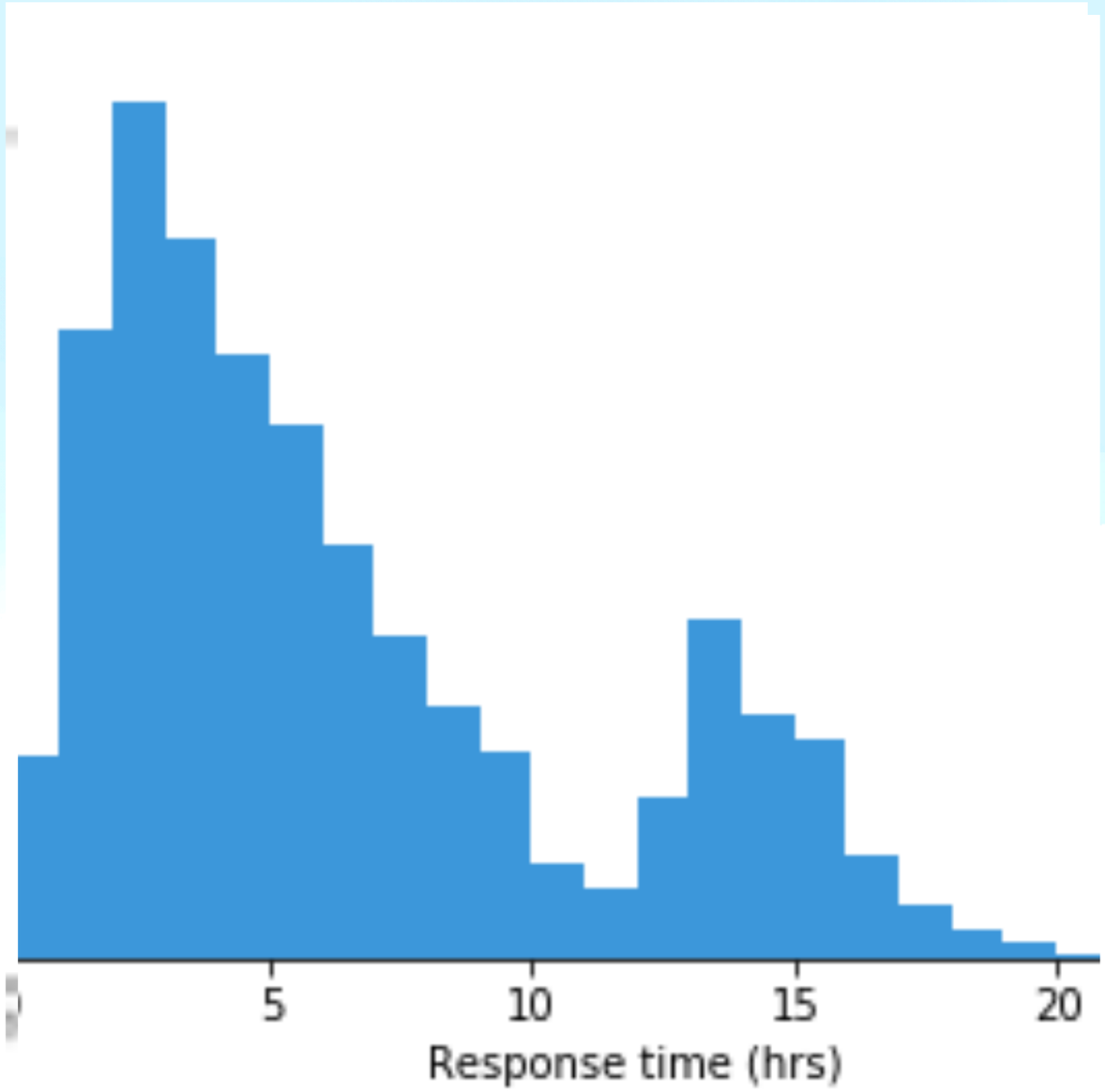
- or fraction of total N,

$$\frac{n_i}{N}$$

- or density

$$\frac{n_i}{N\Delta x}$$

- where the last choice means it will integrate to 1.0.

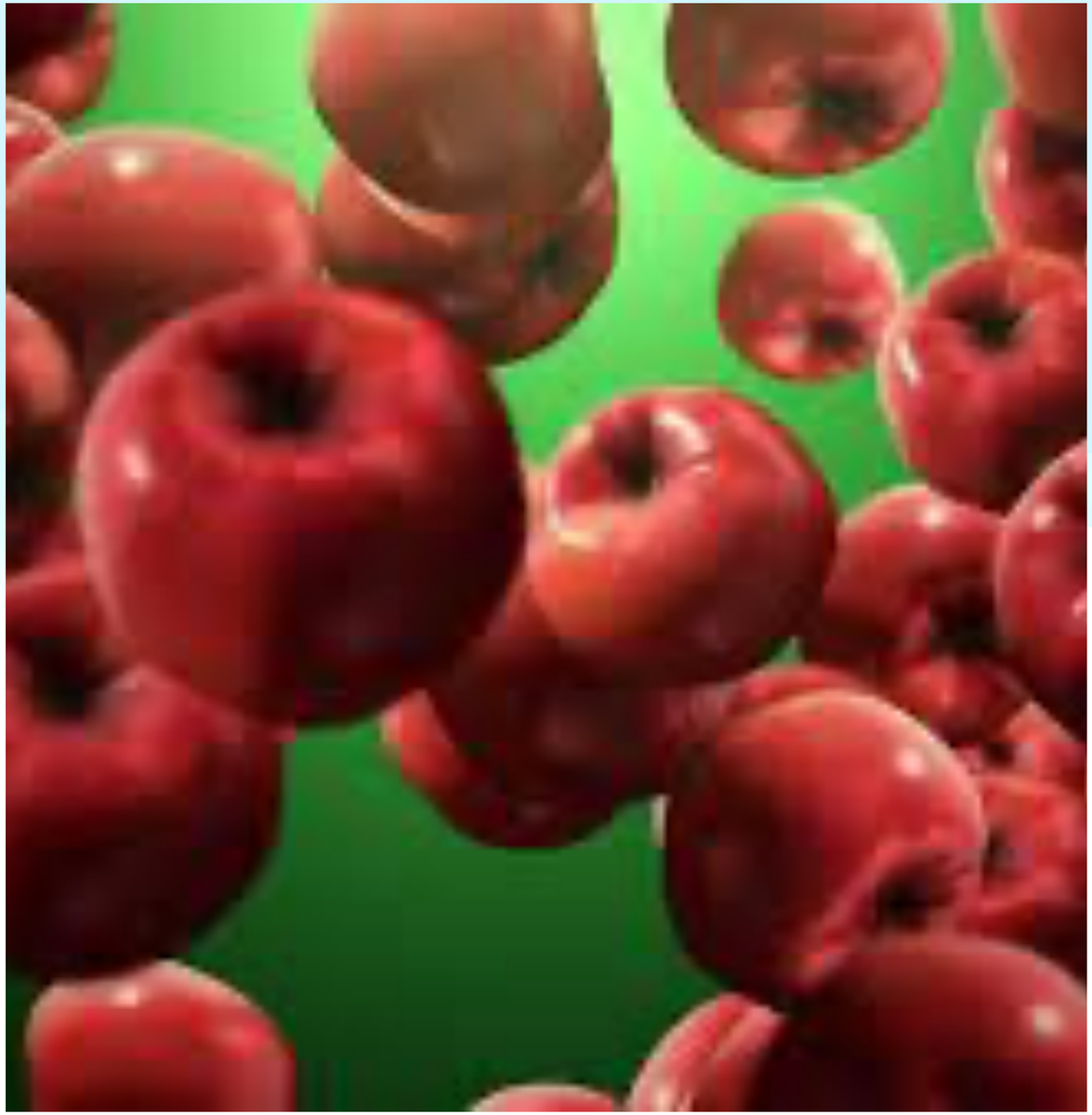


Histogram Code

- Histograms can be made with the hist() function in matplotlib.pylot or the histplot() function in seaborn. It can be useful to control the range over which the histogram is plotted and the number of bins.
  - n,bins,patches = plt.hist(vals,bins=20,range=(0,5))
- However, sometimes it is useful to calculate the histogram using numpy and then plot it later. This can give you more control over how the histogram is plotted or make calculations on the values. The function returns the occruances and the edges of the bins used.
  - hist,bin_edges = np.histogram(vals,bins=20,range=(0,5))
  - plt.stairs(hist,bin_edges)

Populations

- We have been talking about multiple measurements differing because of errors, but they can also differ because you measure different things.
- If we were to weigh a bunch of apples we would not expect them all to have the same mass. We would get a distribution of mass, that represent the population of apples we measured.
- In many ways it is useful to think of measurement errors the same way, that there is a population of results we might get and when we make measurements we sample from those possibilities.



Summary Statistics

central tendency

- The problem with distribution is that they are hard to work with. What to we do with 50 numbers for a measured mass or time?
- Summary statistics allow us to summarize the distribution in a few numbers, but at the price of loosing information.
- The most common summary statistics are the average, median, and mode.
  - Average - add all values and divide by how many $\mu = \frac{1}{N} \sum_{1}^{N-1} x_i$
  - Median - the value where half of the distribution is less than and half is greater than
  - Mode - the most common occurrence in the distribution, better for discrete distributions, continuous distributions have to be binned.

Summary Statstics

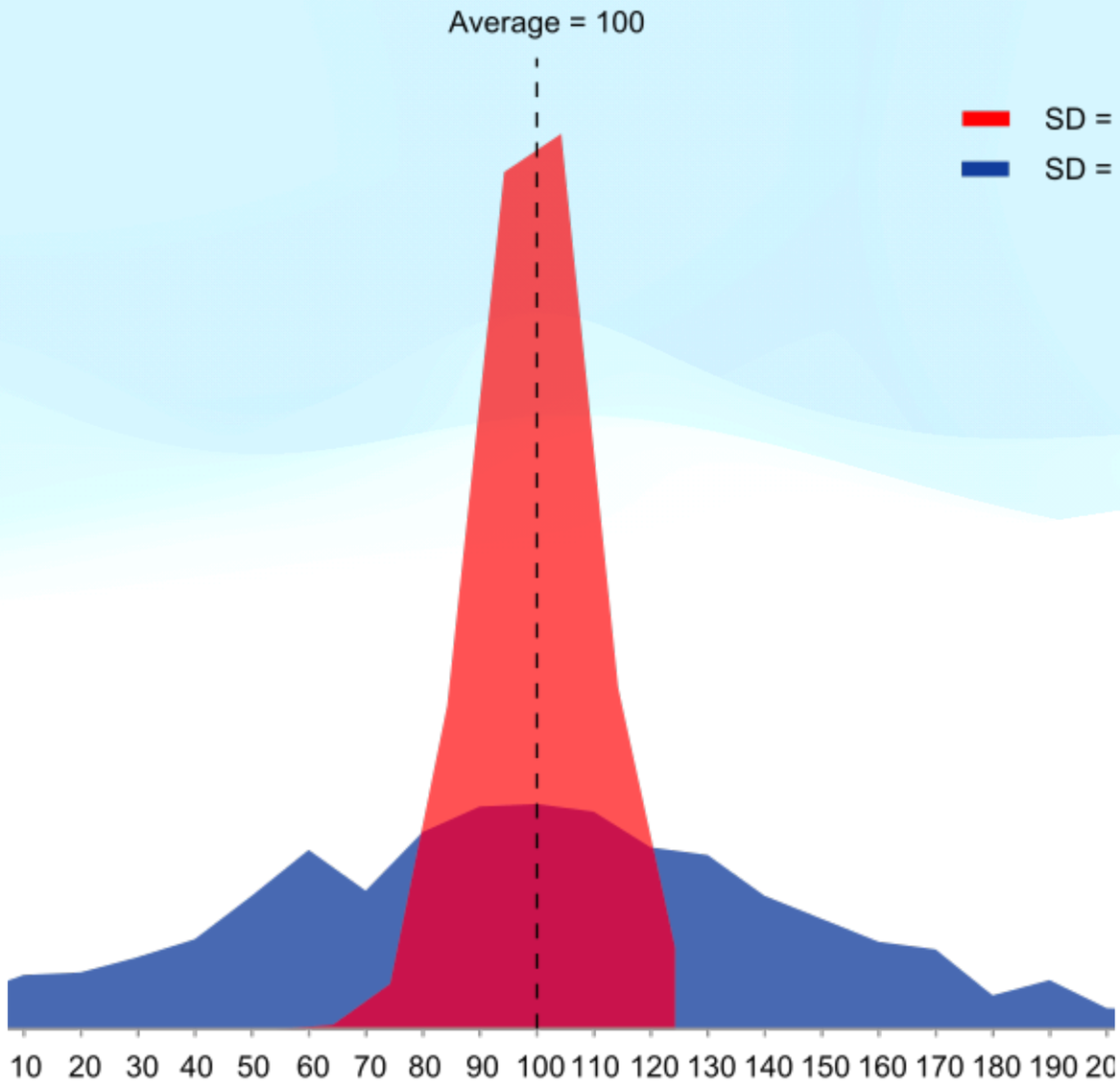
dispersion

- We would also like to know about the width (dispersion) of our distribution.
- This can be captured in the variance

$$\sigma^2 = \frac{1}{N}\sum_N (x_i - \mu)^2$$

- or standar deviation

$$\sigma = \sqrt{\frac{1}{N}\sum_N (x_i - \mu)^2}$$



Summary Statstics

dispersion

- We would also like to know about the width (dispersion) of our distribution.
- This can be captured in the variance

$$\sigma^2 = \frac{1}{N}\sum_N (x_i - \mu)^2$$

- or standar deviation

$$\sigma = \sqrt{\frac{1}{N}\sum_N (x_i - \mu)^2}$$

- The are other measures of the dispersion.
- Interquartile range (IQR), the range from the 25% to 75% of the distribution.
- The range, but this can be dominated by outliers.
- Mean absolute deviation (MAD), this is the median of the absolute value of the residuals.

$$MAD = median(abs(x_i - \mu))$$

Summary Statistics

higher order statistics

- skewness is a measure of the symmetry of a distribution

$$skewness = \frac{\frac{1}{N}\sum_N (x_i - \mu)^3}{[\frac{1}{N}\sum_N (x_i - \mu)^2]^{3/2}}$$

- kurtosis is a measure of the tails of a distribution

$$kurtosis = \frac{\frac{1}{N}\sum_N (x_i - \mu)^4}{[\frac{1}{N}\sum_N (x_i - \mu)^2]^2} - 3$$