Variability
Randomness
Uncertainty

Quantify the variability in the data

\[ \text{Range} : \max \text{ data value} - \min \text{ data value} \]

- not robust in the presence of outliers.

\[ \Rightarrow \text{get rid of the outliers!} \]

Compute quartiles.

1st Quartile: one quarter of the data is less than Q1.

2nd Quartile: median

3rd Quartile: 3/4 of the data is less than Q3.
\[ \frac{1}{4} \text{ of the data is here} \]

\[ \Rightarrow Q_1 \text{ is a data value} \]
A measure of the spread of the data is **inter-quartile range**

\[ Q_3 - Q_1 \]

→ robust to outliers
→ good for non-symmetric distributions

Quartiles are just specific percentiles.

- given a percentile, e.g., 10th percentile:
  count up through the sorted data until you have seen 10% of the data. The data value you have reached is the 10th percentile value.

- given a data value, its percentile is given by the fraction of the data that has smaller values.
Another measure of the spread, (that is more convenient mathematically) is the **standard deviation**.

what is the "average size" of a set of numbers?

leaves thickness 33 36 36 37 39 40

"size" is around 36 or 37

what about: 0 5 -8 7 -3

average is 0.2.

most of the numbers are quite different from 0.2.

→ need to make all the numbers positive.

1/ square all the numbers 0 25 64 49 9.

2/ compute average of the squared values \[
\frac{0+25+64+49+9}{5} = 29.4.
\]

3/ take the square root \[
\sqrt{29.4} = 5.4
\]
This is called the Root Mean Square Value (RMS) (has nicer mathematical properties than just dropping the negative signs).

Measure of the spread:

- RMS deviation about the mean
  - on average, how far away are the data from the mean.

![Histograms]

Small SD  

![Histograms]

Large SD

Computing the SD.

20, 10, 15, 15 ← data.
1. Compute the mean

\[ \frac{20 + 10 + 15 + 15}{4} = \frac{60}{4} = 15 \]

2. Compute the deviation of each data value from the mean.

- \(20 \text{-} 15 = 5\)
- \(10 \text{-} 15 = -5\)
- \(15 \text{-} 15 = 0\)
- \(15 \text{-} 15 = 0\)

3. Calculate the RMS of the deviations

\[ \sqrt{\frac{(5)^2 + (-5)^2 + 0^2 + 0^2}{4}} = \sqrt{\frac{25 + 25 + 0 + 0}{4}} = \sqrt{\frac{50}{4}} = \sqrt{12.5} = 3.5 \]

S.D. = 3.5

A measure of the spread about the mean.
Note: Make sure you're using the correct SD button on your calculator.
- we're dividing by N, not (N-1)

Most observations are within \( \pm 1 \) SD of the mean.

Few are more than 2SD away from the mean.

68\% of entries (\( \pm 2 \) in 3) are within 1 SD of mean.

95\% \hspace{1cm} (19 \text{ in} \ 20) \hspace{1cm} 2 \text{SD}

This holds best for symmetric data with a symmetric distribution, but is often ok even for data with non-symmetric distributions.
About 68% of the data is within ±1σ of the mean.
Standard Normal curve

\[ \uparrow \]
mean = 0
median = 0  (symmetric)
SD = 1.
68% of the area under the curve is in the range 
\[-1 \leq x \leq 1\]
do not expect to see data values more than 4 SD from the mean.

95%.
Normal Approximation for Data.

Very many (but by no means all) data histograms can be approximated by the Normal Curve.

\[ y = \frac{100\%}{\sqrt{2\pi}} e^{-x^2/2} \]

For Normal curve:

- Between -1 and +1, area under curve = 68%.
- Between -2 and +2, area under curve = 95%.
- Between -3 and +3, area under curve = 99.7%.

X-axis is in standard units. What does that mean?

Standard units are the number of SD away from the mean.
Example: Women in a survey had mean height 63.5 in
SD 3 in

One particular woman was 69.5 in tall.

correct to standard units:

69.5 in 6 in more than average
6 in is twice SD

so 69.5 in is \( \frac{2}{3} \) in standard units.

Another woman was 62 in tall.

correct to standard units \( \frac{62 - 63.5}{3} = \frac{-1.5}{3} = -0.5 \)
units for the
normal curve.

standard units

units for the
histogram.
N 68% of data are within 1 SD of mean

95%
Converting into standard units allows us to use a single Normal curve.

To approximate the % of entries in an interval in data scale

(e.g. % of leaves with water fraction < 0.65)

→ convert that interval into standard units.
→ find the area under the normal curve in this interval.

( area under normal curve to left of z = -0.8 ).

- We find areas under the normal curve by using the table in the back of the textbook.

- the table has areas for symmetric intervals.
area is 95.45%.

By symmetry, one of the red areas is \( \frac{4.2^2}{2} \approx 21\% \).

Two red areas together are 100 - 58 = 42\%.

This is the area in the table. = 58\%
By combining
- central area from the table

with knowledge that
- the curve is symmetric
- the total area is 100%

we can compute the area underneath the curve for any interval.
\[
= \frac{1}{2} \times 95\%
+ \frac{1}{2} \times 68\%.
\]

\[
= 82\%.
\]

If we would expect about 8% of our data set to have values more than 1.4 SD above the mean.
Normal Approximation for Data.

Key: convert the region of data space that we're interested in into standard units, using the mean ± SD of the data to do so.

Use the standard normal curve to obtain the % of the data that falls into the region.

Example: the heights of a group of men had mean 69 in

SD 3 in

What % of the men had heights between 63 and 72 in?

Assuming that the histogram of the heights follows the Normal curve.
data space
(height in inches)

standard units.

\[
\frac{95}{2} + \frac{68}{2} = \frac{82}{2}
\]
Women's heights mean 63.5
SD 3.0.

% with heights > 59 in

59 in standard units \( \frac{59 - 63.5}{3} = -1.5 \)

-1.5

93.5%

87%

-1.5 1.5

Together have 22%.

6.5

100 - 6.5

\( \pm 93.5\% \)