and gathers the following data:

age	weight
0	8.2
1	7.8
2	10.1
3	11.3
4	12.5
5	15.1
6	16.1

Present a scatter plot of these data. Do you think there is a connection between age and weight?

Video Assignment.

View the following program(s) from the series *Against All Odds*:

Program	*Title*
One	*What is Statistics?*
Two	*Picturing Distributions*

II. Descriptive Statistics: Means and Variances

Given a collection of data, the *mean* is just the numerical average. Conceptually there are two kinds of data:

- census data for the entire population; and
- sample data for a part of the population.

Thus there are two kinds of means:

- population means; and
- sample means.

The *population mean* is denoted by the Greek letter $\boldsymbol{\mu}$ (mu). The *sample mean* is denoted by the letter $\bar{\boldsymbol{x}}$ (x bar). Fortunately, whether you have sample data or census data, there is only one formula for the mean:

$$\text{mean} = \frac{1}{\boldsymbol{n}} \times (\boldsymbol{x}_1 + \boldsymbol{x}_2 + \cdots + \boldsymbol{x}_{\boldsymbol{n}})$$

where $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_{\boldsymbol{n}}$ are the data points and $\boldsymbol{n}$ is the number of data points.

Since the above sum is somewhat cumbersome to write, there is a shorthand notation called "summation notation." In particular, it is customary to write

$$\begin{aligned} \bar{\boldsymbol{x}} &= \frac{1}{\boldsymbol{n}} \times (\boldsymbol{x}_1 + \boldsymbol{x}_2 + \cdots + \boldsymbol{x}_{\boldsymbol{n}}) \\ &= \frac{1}{\boldsymbol{n}} \sum_{\boldsymbol{i}=1}^{\boldsymbol{n}} \boldsymbol{x}_{\boldsymbol{i}}. \end{aligned}$$

The last line is just shorthand for the first line.

Example 1. Given the following sample data, find the sample mean.

12	15	18	16
19	12	22	2

Solution. In this problem, there are 8 data points, so $\boldsymbol{n} = 8$. Then

$$\begin{aligned}\bar{\boldsymbol{x}} &= \frac{1}{\boldsymbol{n}} \sum_{\boldsymbol{i}=1}^{\boldsymbol{n}} \boldsymbol{x_i} \\ &= \frac{1}{8} (12 + 15 + 18 + 16 + 19 + 12 + 22 + 2) \\ &= \frac{116}{8} \\ &= 14.5\end{aligned}$$

and thus the sample mean $\bar{\boldsymbol{x}}$ is 14.5.

Example 2. Given the following census data, find the population mean.

14	15	16	13
15	18	13	12

Solution. Once again there are eight data points, so

$$\begin{aligned}\boldsymbol{\mu} &= \frac{1}{\boldsymbol{n}} \sum_{\boldsymbol{i}=1}^{\boldsymbol{n}} \boldsymbol{x_i} \\ &= \frac{1}{8} (14 + 15 + 16 + 13 + 15 + 18 + 13 + 12) \\ &= \frac{116}{8} \\ &= 14.5\end{aligned}$$

and thus the population mean $\bar{\boldsymbol{x}}$ is 14.5.

Note that we get the same mean in both examples, but that the data themselves are qualitatively different. In particular, in example one the scores range from 2 to 22 while in example two the scores range only from 12 to 18. In example two, the data points appear to be "nearer to" the sample mean than in example one; another way of saying this is that the data points in example one have more variability than those in example two. The variance is one way to measure the relative variability of the two samples.

One way to estimate the variability in each sample might be to compute ($\boldsymbol{x_i}$−mean) for each observation $\boldsymbol{x_i}$ and then average these differences. The relevant calculations are in the tables below.

x_i	$x_i - \bar{x}$
12	-2.5
15	0.5
18	3.5
16	1.5
19	4.5
12	-2.5
22	7.5
2	-12.5
116	0

x_i	$x_i - \mu$
14	-0.5
15	0.5
16	1.5
13	-1.5
15	0.5
18	-3.5
13	-1.5
12	-2.5
116	0

In each table, the positive and negative terms in the difference column exactly balance out to zero. This makes sense if you examine what happens when you add up the difference column. For example, for the numbers in example one:

$$\begin{aligned}\sum_{i=1}^{8}(x_i - \bar{x}) &= (12 - 14.5 + 15 - 14.5 + 18 - 14.5 + 16 - 14.5 + 19 - \cdots \\ &\quad \cdots 14.5 + 12 - 14.5 + 22 - 14.5 + 2 - 14.5) \\ &= \left(\underbrace{12 + 15 + 18 + 16 + 19 + 12 + 22 + 2}_{8\times\bar{x}} - \cdots \right. \\ &\quad \left. \cdots \underbrace{14.5 - 14.5 - 14.5 - 14.5 - 14.5 - 14.5 - 14.5 - 14.5}_{8\times\bar{x}} \right) \\ &= 0\end{aligned}$$

In exactly the same way, the difference column for the numbers in example two sums to zero.

There are several remedies which might be applied in order to avoid this cancellation. The most common is to *square* the differences so that all of the negative signs go away.

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
12	-2.5	6.25
15	0.5	0.25
18	3.5	12.25
16	1.5	2.25
19	4.5	20.25
12	-2.5	6.25
22	7.5	56.25
2	-12.5	156.25
116	0	260

x_i	$x_i - \mu$	$(x_i - \mu)^2$
14	-0.5	0.25
15	0.5	0.25
16	1.5	2.25
13	-1.5	2.25
15	0.5	0.25
18	3.5	12.25
13	-1.5	2.25
12	-2.5	6.25
116	0	26

Thus, in the first example the *average* "mean square" error is $\frac{1}{8} \times 260 = 32.5$ while in the second example the average "mean square" error is $\frac{1}{8} \times 26 = 3.25$. This confirms our previous intuition that the first sample was more variable than the second.

The quantity computed above is called the *variance*. As with means, there are two kinds of variances, depending on whether or not you started with census or sample data. Unfortunately, for technically reasons beyond the scope of this course, there are actually *two* formulae for the variance depending on whether you started with sample data or with census data.

If you start with *census* data, then you are computing the *population variance* denoted by the Greek letter $\boldsymbol{\sigma}^2$ (lower case sigma squared) and the formula is

$$\boldsymbol{\sigma}^2 = \frac{1}{\boldsymbol{n}} \sum_{\boldsymbol{i}=1}^{\boldsymbol{n}} (\boldsymbol{x_i} - \bar{\boldsymbol{x}})^2.$$

In particular, we had census data for example two, so the population variance is the number we computed (the average of the squared differences): 3.25.

If you start with *sample* data then you are computing the *sample* variance denoted by the symbol $\boldsymbol{s}^2$ and the formula is

$$\boldsymbol{s}^2 = \frac{1}{\boldsymbol{n}-1} \sum_{\boldsymbol{i}=1}^{\boldsymbol{n}} (\boldsymbol{x_i} - \bar{\boldsymbol{x}})^2.$$

Since we had sample data in example one, the sample variance is

$$\boldsymbol{s}^2 = \frac{1}{8-1} 260 = 37.143.$$

The