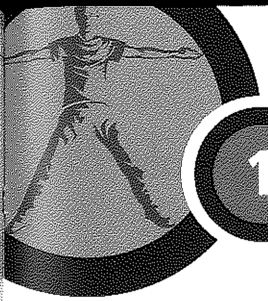
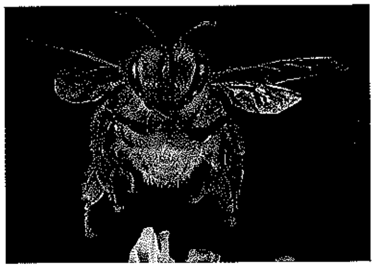


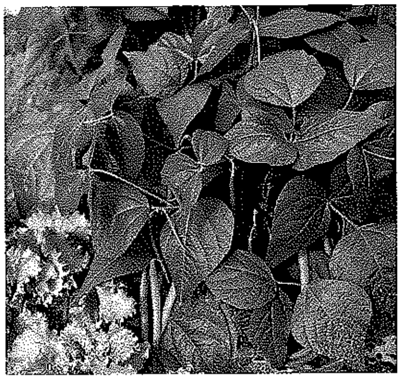
1 Statistical analysis

Introduction



This is an Africanized honey bee (AHB). AHBs have spread to the USA from Brazil. They are now in competition with the local bee population, which are European honey bees (EHBs). EHBs were brought to America by European colonists in the 1600s. AHBs are now out-competing EHBs in areas the former invade.

This mixed oak forest is a mature stage in the development of plant communities surrounding bodies of water. Studying the growth rate of the trees such as maple, beech, oak and hickory gives us evidence of the health of these plant communities.



This is the common bean plant used by many students in their classrooms. Bean plants grow in about 30 days under banks of artificial lights. Seeds are easy to obtain. Germinated seeds can be placed in paper cups with sterilized soil. Many factors can be tested to determine whether or not they affect the growth of the bean plants.

In this chapter, you will learn how scientists analyse the evidence they collect when they perform experiments. You will be designing your own experiments, so this information will be very useful to you. You will be learning about:

- means;
- error bars;
- t-test;
- standard deviation;
- significant difference;
- causation and correlation.

Have your calculator by you to practise calculations for standard deviation and t-test so that you can use these methods of analysing data when you do your own experiments.

1.1 Statistics

Assessment statements

1.1.1 State that error bars are a graphical representation of the variability of data.

1.1.2 Calculate the mean and standard deviation of a set of values.

1.1.3 State that the term standard deviation is used to summarize the spread of values around the mean, and that 68% of values fall within one standard deviation of the mean.

1.1.4 Explain how the standard deviation is useful for comparing the means and spread of data between two or more samples.

1.1.5 Deduce the significance of the difference between two sets of data using calculated values for *t* and the appropriate tables.

1.1.6 Explain that the existence of a correlation does not establish that there is a causal relationship between two variables.

Reasons for using statistics

Statistics can be used to describe conditions which exist in countries around the world. The data can heighten our awareness of global issues. To learn about United Nations poverty facts and consumption statistics, visit heinemann.co.uk/hotlinks, insert the express code 4273P and click on Weblinks 1.1a and 1.1b.

Biology examines the world in which we live. Plants and animals, bacteria and viruses all interact with one another and the environment. In order to examine the relationships of living things to their environments and each other, biologists use the scientific method when designing experiments. The first step in the scientific method is to make observations. In science, observations result in the collection of measurable data. For example: What is the height of bean plants growing in sunlight compared to the height of bean plants growing in the shade? Do their heights differ? Do different species of bean plants have varying responses to sunlight and shade? After we have observed, we then decide which of these questions to answer. Assume we want to answer the question, 'Will the bean plant, *Phaseolus vulgaris*, grow taller in sunlight or in the shade?' We must design an experiment which can try to answer this question.

How many bean plants should we use in order to answer our question? Obviously, we cannot measure every bean plant that exists. We cannot even realistically set up thousands and thousands of bean plants and take the time to measure their height. Time, money, and people available to do the science are all factors which determine how many bean plants will be in the experiment. We must use samples of bean plants which represent the population of all bean plants. If we are growing the bean plants, we must plant enough seeds to get a representative sample.

The World Health Organization (WHO) website has statistical information on infectious diseases. Over the next year, the system aims to provide a single point of access to data, reports and documents on:

- major diseases of poverty including malaria, HIV/AIDS, tuberculosis;
- diseases on their way towards eradication and elimination such as guinea worm, leprosy, lymphatic filariasis;
- epidemic-prone and emerging infections such as meningitis, cholera, yellow fever and anti-infective drug resistance.

Statistics is a branch of mathematics which allows us to sample small portions from habitats, communities, or biological populations, and draw conclusions about the larger population. Statistics mathematically measures the differences and relationships between sets of data. Using statistics, we can take a small population of bean plants grown in sunlight and compare it to a small population of bean plants grown in the shade. We can then mathematically determine the differences between the heights of these bean plants. Depending on the sample size that we choose, we can draw conclusions with a certain level of confidence. Based on a statistical test, we may be able to be 95% certain that bean plants grown in sunlight will be taller than bean plants grown in the shade. We may even be able to say that we are 99% certain, but nothing is 100% certain in science.

Some questions

- In order to determine the chances that men over 40 years of age will have a heart attack, must we look at the number of heart attacks which occurred in *all* men over 40 years of age?

  No, we can take a representative sample of men over 40 years of age and, using statistics, determine the risk of men over 40 years of age having a heart attack.

- If we compare the heights of boys who are 16 years of age living in Britain with the heights of boys 16 years of age who are living in the US, must we have the heights of *all* of the boys from the US and all of the boys from Britain?

  No, we can take representative sample of 16 year old boys from the US and a sample of boys from Britain and use mathematics to determine if there is a statistically significant difference in their heights.

- How confident can we be that the difference between the heights of boys in the US and the heights of boys in the UK is significantly different based on a representative sample?

  We can be up to 99% confident in the difference. The only way to be 100% confident is if we measured the height of every boy!

For more information and access to the WHO website, visit heinemann.co.uk/hotlinks, insert the express code 4273P and click on Weblink 1.2.

For more statistics on the AIDS epidemic, visit heinemann.co.uk/hotlinks, insert the express code 4273P and click on Weblink 1.3.

Mean, range, standard deviation and error bars

Statistics analyses data using the following terms:

- mean;
- range;
- standard deviation;
- error bars.

Mean

The mean is an average of data points. For example, suppose the height of bean plants grown in sunlight is measured in centimetres at 10 days after planting. The heights are 10, 11, 12, 9, 8 and 7 centimetres. The sum of the heights is 57 centimetres. Divide 57 by 6 to find the mean (average). The mean is 9.5 centimetres. The mean is the central tendency of the data.

Range

The range is the measure of the spread of data. It is the difference between the largest and the smallest observed values. In our example, the range is 12 – 7 = 5. The range for this data set is 5 centimetres. If one data point were unusually large or unusually small, this very large or small data point would have a great effect on the range. Such very large or very small data points are called outliers. In our sample there is no outlier.

Standard deviation

The standard deviation (SD) is a measure of how the individual observations of a data set are dispersed or spread out around the mean. Standard deviation is determined by a mathematical formula which is programmed into your calculator. You can calculate the standard deviation of a data set by using the SD function of a graphic display or scientific calculator.

Error bars

Error bars are a graphical representation of the variability of data. *Error bars can be used to show either the range of data or the standard deviation on a graph.* Notice the error bars representing standard deviation on the histogram in Figure 1.1 and the line graph in Figure 1.2. The value of the standard deviation above the mean is shown extending above the top of each bar of the histogram and the same standard deviation below the mean is shown extending below the top of each bar

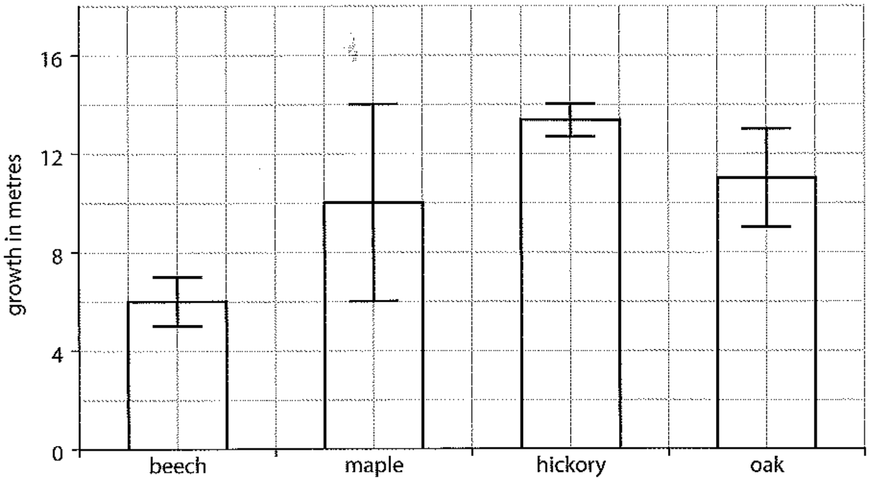


Figure 1.1 Rate of tree growth on the Oak–Hickory Dune 2004–05. Values are represented as mean ±1SD from 25 trees per species.

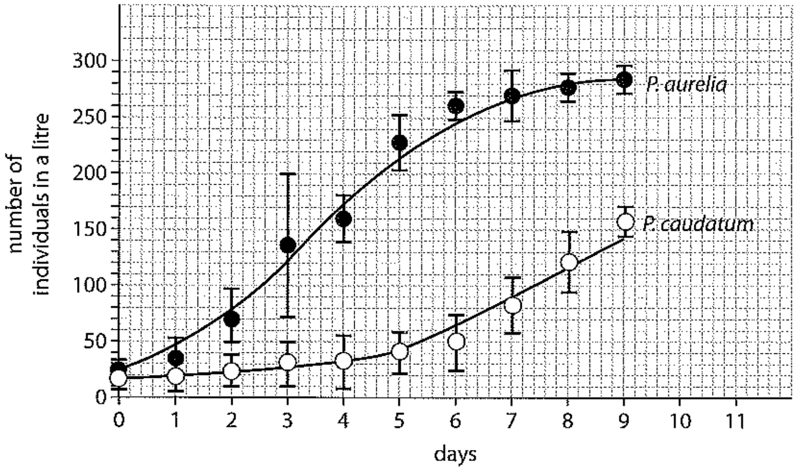


Figure 1.2 Mean population density (±1SD) of two species of *Paramecium* grown in solution.

of the histogram. Since each bar represents the mean of the data, the standard deviation for each type of tree will be different, but the values extending above and below one bar will be the same. The same is true for the line graph. Since each point on the graph represents the mean data for each day, the bars extending above and below the data point are the standard deviations above and below the mean.

For example, let us look at Figure 1.1. Notice that the error bar on the beech tree growth bar extends up 1 metre above the top of the bar and below the top of the bar by 1 metre. Thus, the SD for beech trees is 1 metre. Since the top of the bar is the mean of 6 then we can say that the average growth of beech trees is 6 metres ± the standard deviation of 1 metre. Look at the bar for maple trees. The top of the bar or mean is at 10 metres. Notice that that the error bar extending above the mean is 4 metres and it also extends below the mean by 4 metres. Thus, the average growth for maple trees is 10 metres ± the standard deviation of 4 metres. Since the standard deviation of the beech tree is smaller than the standard deviation of maple trees, we are more confident of the accuracy of the beech tree data. The maple tree data is more variable.

Now let us look at the line graph of Figure 1.2. The caption states that the dots are representing the mean population density ±1 standard deviation. Find the number of individuals for *P. aurelia* on day 3. As you can see, the average number of individuals is 135 and the standard deviation is ±65 individuals. However, for *P. caudatum* on day 3 the SD is much less. The average is 30 individuals and the SD is ±20 individuals. Since the standard deviation of *P. caudatum* is less than for *P. aurelia*, we are more confident of the data for *P. caudatum*. The data for *P. aurelia* is more variable.

Standard deviation

We use standard deviation to summarize the spread of values around the mean and to compare the means and spread of data between two or more samples.

Summarizing the spread of values around the mean

In a normal distribution, about 68% of all values lie within ±1 standard deviation of the mean. This rises to about 95% for ±2 standard deviations from the mean.

To help understand this difficult concept, let's look back to the bean plants growing in sunlight and shade. First, the bean plants in the sunlight: suppose our sample is 100 bean plants. Of that 100 plants, you might guess that a few will be very short (maybe the soil they are in is slightly sandier). A few may be much taller than the rest (possibly the soil they are in holds more water). However, all we can measure is the height of all the bean plants growing in the sunlight. If we then plot a graph of the heights, the graph is likely to be similar to a bell curve (see Figure 1.3). In this graph, the number of bean plants is plotted on the *y* axis and the heights ranging from short to medium to tall are plotted on the *x* axis.

Many data sets do not have a distribution which is this perfect. Sometimes, the bell-shape is very flat. This indicates that the data is spread out widely from the mean. In some cases, the bell-shape is very tall and narrow. This shows the data is very close to the mean and not spread out.

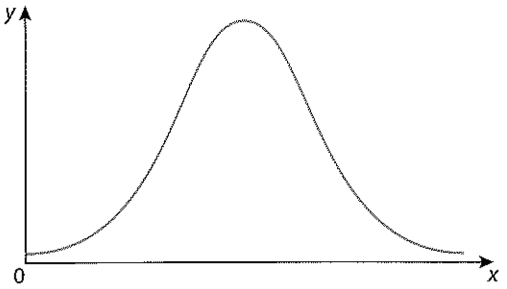


Figure 1.3 This graph shows a bell curve.

The standard deviation tells us how tightly the data points are clustered around the mean. When the data points are clustered together, the standard deviation is small; when they are spread apart, the standard deviation is large. Calculating the standard deviation of a data set is easily done on your calculator.

Look at Figure 1.4. This graph of normal distribution may help you understand what standard deviation really means. The dotted area represents one standard deviation in either direction from the mean. About 68% of the data in this graph is located in the dotted area. Thus, we say that for normally distributed data, 68% of all values lie within ±1 standard deviation from the mean. Two standard deviations from the mean (the dotted and the cross-hatched areas) contain about 95% of the data. If this bell curve were flatter, the standard deviation would have to be larger to account for the 68% or 95% of the data set. Now you can see why standard deviation tells you how widespread your data points are from the mean of the data set.

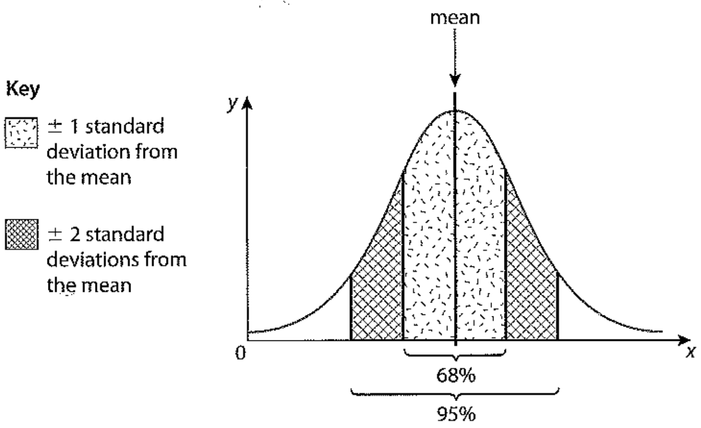


Figure 1.4 This graph shows a normal distribution.

For directions on how to calculate standard deviation with a TI-86 calculator, visit heinemann.co.uk/hotlinks, insert the express code 4273P and click on Weblink 1.4a.

If you have a TI-83 calculator, visit heinemann.co.uk/hotlinks, insert the express code 4273P and click on Weblink 1.4b.

How is this useful? For one thing, it tells you how many extremes are in the data. If there are many extremes, the standard deviation will be large; with few extremes the standard deviation will be small.

Question

- What is the shape of the graph of a normal distribution of data points?

  The shape of the graph is a bell curve.

Worked example 1.1

If there are 100 bean plants represented by the bell curve, how many will be within one standard deviation of the mean?

Solution

68 bean plants will be within one standard deviation of the mean since 68 out of 100 is 68%.

Comparing the means and spread of data between two or more samples

Remember that in statistics we make inferences about a whole population based on just a sample of the population. Let's continue using our example of bean plants growing in the sunlight and shade to determine how standard deviation is useful for comparing the means and the spread of data between these two samples. Here are the raw data sets for bean plants grown in sunlight and in shade.

Height of bean plants in the sunlight in centimetres ±0.1 cm	Height of bean plants in the shade in centimetres ±0.1 cm
124	131
120	60
153	160
98	212
123	117
142	65
156	155
128	160
139	145
117	95
Total 1300	Total 1300

First, we determine the mean for each sample. Since each sample contains 10 plants, we can divide the total by 10 in each case. The resulting mean is 130.0 centimetres for each condition.

Of course, that is not the end of the analysis. Can you see there are large differences between the two sets of data? The height of the bean plants in the shade is much more variable than that of the bean plants in the sunlight. The means of each data set are the same, but the variation is not the same. This suggests that other factors may be influencing growth in addition to sunlight and shade.

How can we mathematically quantify the variation that we have observed? Fortunately, your calculator has a function that will do this for you. All you have to do is input the raw data. For practice, find the standard deviation of each raw data set above before you read on.

The standard deviation of the bean plants growing in sunlight is 17.68 centimetres while the standard deviation of the bean plants growing in the shade is 47.02 centimetres. Looking at the means alone, it appears that there is no difference between the two sets of bean plants. However, the high standard deviation of the bean plants grown in the shade indicates a very wide spread of data around the mean. The wide variation in this data set makes us question the experimental design. Is it possible that the plants in the shade are also growing in several different types of soil? What is causing this wide variation in data? This is why it is important to calculate the standard deviation in addition to the mean of a data set. If we looked at only the means, we would not recognize the variability of data seen in the shade-grown bean plants.

W To use an online calculator to do the t-test go to: heinemann.co.uk/hotlinks, insert the express code 4273P and click on Weblink 1.5a or 1.5b.

Worked example 1.2

If all the data values are equal, such as 7, 7, 7, 7, what is the standard deviation of this set of four data points?

Solution

Zero! If all the data points are the same, there is no deviation from the mean.

Worked example 1.3

If the daily temperatures of a city A range from 10 °C to 30 °C for one month, the mean temperature may be 20 °C. Another city B may also have a mean temperature of 20 °C for the same month. However, the range of city B is only 15 °C to 25 °C.

Which city has a temperature with a higher standard deviation?

Which city can give a more accurate prediction of the weather and why?

Solution

City A has a higher standard deviation.

City B since it has a very narrow range or temperature or a very low standard deviation.

Significant difference between two data sets using the t-test

In order to determine whether or not the difference between two sets of data is a significant (real) difference, the t-test is commonly used. The t-test compares two sets of data, for example heights of bean plants grown in the sunlight and heights of bean plants grown in the shade. Look at the bottom of the table of *t* values (page 8), and you will see the probability (p) that chance alone could make a difference. If $p = 0.50$, we see the difference is due to chance 50% of the time. This is not a significant difference in statistics. However, if you reach $p = 0.05$, the probability that the difference is due to chance is only 5%. That means that there is a 95% chance that the difference is due (in our bean example) to one set of the bean plants being in the sunlight. A 95% chance is a significant difference in statistics. Statisticians are never completely certain but they like to be at least 95% certain of their findings before drawing conclusions.

Let's look at the t-table in a new way. Look again at the bottom of the table. Using our example of bean plants, if $p = 0.50$, we see that the Sun causes the difference between the two groups for 50% of the time. That means that for the other 50%

Table of t values

Degrees of freedom	*t* values					
1	1.00	3.08	6.31	12.71	63.66	636.62
2	0.82	1.89	2.92	4.30	9.93	31.60
3	0.77	1.64	2.35	3.18	5.84	12.92
4	0.74	1.53	2.13	2.78	4.60	8.61
5	0.73	1.48	2.02	2.57	4.03	6.87
6	0.72	1.44	1.94	2.45	3.71	5.96
7	0.71	1.42	1.90	2.37	3.50	5.41
8	0.71	1.40	1.86	2.31	3.367	5.04
9	0.70	1.38	1.83	2.26	3.25	4.78
10	0.70	1.37	1.81	2.23	3.17	4.590
11	0.70	1.36	1.80	2.20	3.11	4.44
12	0.70	1.36	1.78	2.18	3.06	4.32
13	0.69	1.35	1.77	2.16	3.01	4.22
14	0.69	1.35	1.76	2.15	2.98	4.14
15	0.69	1.34	1.75	2.13	2.95	4.07
16	0.69	1.34	1.75	2.12	2.92	4.02
17	0.69	1.33	1.74	2.11	2.90	3.97
18	0.69	1.33	1.73	2.10	2.88	3.92
19	0.69	1.33	1.73	2.09	2.86	3.88
20	0.69	1.33	1.73	2.09	2.85	3.85
21	0.69	1.32	1.72	2.08	2.83	3.82
22	0.69	1.32	1.72	2.07	2.82	3.79
24	0.69	1.32	1.71	2.06	2.80	3.75
26	0.68	1.32	1.71	2.06	2.78	3.71
28	0.68	1.31	1.70	2.05	2.76	3.67
30	0.68	1.31	1.70	2.04	2.75	3.65
35	0.68	1.31	1.69	2.03	2.72	3.59
40	0.68	1.30	1.68	2.02	2.70	3.55
45	0.68	1.30	1.68	2.01	2.70	3.52
50	0.68	1.30	1.68	2.01	2.68	3.50
60	0.68	1.30	1.67	2.00	2.66	3.46
70	0.68	1.29	1.67	1.99	2.65	3.44
80	0.68	1.29	1.66	1.99	2.64	3.42
90	0.68	1.29	1.66	1.99	2.63	3.40
100	0.68	1.29	1.66	1.99	2.63	3.39
Probability (p) that chance alone could produce the difference	0.50 (50%)	0.20 (20%)	0.10 (10%)	0.05 (5%)	0.01 (1%)	0.001 (0.1%)

of the time the Sun did *not* make a difference. It makes us doubt the importance of the Sun on bean plants. We are only 50% confident that the Sun makes the bean plants grow taller.

What if we reach a p value of 0.05? This means that the probability that chance alone can cause the difference is only 5%. Thus, we are 95% confident that the Sun is causing the difference in the height of bean plants. A 95% confidence level is an acceptable level of confidence in science.

When comparing two groups of data, we use the mean, standard deviation and sample size to calculate the value of *t*. When given a calculated value of *t*, you can use a table of *t* values. First, look in the left-hand column headed 'Degrees of freedom', then across to the given *t* value. The degrees of freedom are the sum of sample sizes of each of the two groups minus 2.

If the degree of freedom is 9, and if the given value of *t* is 2.60, the table indicates that the *t* value is just greater than 2.26. Looking down at the bottom of the table, you will see that the probability that chance alone could produce the result is only 5% (0.05). This means that there is a 95% chance that the difference is significant.

Worked example 1.4

Compare two groups of barnacles living on a rocky shore. Measure the width of their shells to see if a significant size difference is found depending on how close they live to the water. One group lives between 0 and 10 metres from the water level. The second group lives between 10 and 20 metres above the water level.

Measurement was taken of the width of the shells in millimetres. 15 shells were measured from each group. The mean of the group closer to the water indicates that living closer to the water causes the barnacles to have a larger shell. If the value of *t* is 2.25, is that a significant difference?

Solution

The degree of freedom is 28 (15 + 15 – 2 = 28). 2.25 is just above 2.05.

Referring to the bottom of this column in the table, p = 0.05 so the probability that chance alone could produce that result is only 5%.

The confidence level is 95%. We are 95% confident that the difference between the barnacles is significant. Barnacles living nearer the water have a significantly larger shell than those living 10 metres or more away from the water.

Worked example 1.5

The heights of 16-year-old girls from the UK and the US were compared. The means indicate that the British girls are taller than the US girls. The sample size from each group was 50 girls from each country. Using the formula for *t* which includes the means and standard deviations of the heights from each group, the calculated *t* was found to be 2.00.

What are the degrees of freedom used to determine the probability that the differences between the two groups are due to chance?

Using the given *t* value of 2.00 with your calculated degrees of freedom, what is the probability that chance alone can produce a difference in the heights of these girls?

How confident are we that the British girls are taller than the US girls based on this sample size?

Solution

The degrees of freedom are 50 + 50 – 2 = 98. As the t-table does not show 98, we use the closest number which is 100.

At the bottom of the t-table you will find the probability is 0.05 or 5%. Thus, there is only a 5% chance that the difference is due to chance alone.

We are 95% confident that the difference is a real difference and not just due to chance.

Correlation does not mean causation

We make observations all the time about the living world around us. We might notice, for example, that our bean plants wilt when the soil is dry. This is a simple observation. We might do an experiment to see if watering the bean plants prevents wilting. Observing that wilting occurs when the soil is dry is a simple correlation, but the experiment gives us evidence that the lack of water is the cause of the wilting. Experiments provide a test which shows cause. Observations without an experiment can only show a correlation.

Africanized honey bees

The story of Africanized honey bees (AHBs) invading the USA includes an interesting correlation. In 1990, a honey bee swarm was found outside a small town in southern Texas. They were identified as AHBs. These bees were brought from Africa to Brazil in the 1950s, in the hope of breeding a bee adapted to the South American tropical climate. But by 1990, they had spread to the southern US. Scientists predicted that AHBs would invade all the southern states of the US, but this hasn't happened. Look at Figure 1.5: the bees have remained in the southwest states (area shaded in yellow) and have not travelled to the south-eastern states. The edge of the areas shaded in yellow coincides with the point at which there is an annual rainfall of 137.5cm (55 inches) *spread evenly throughout the year.* This level of *year-round wetness* seems to be a barrier to the movement of the bees and they do not move into such areas.

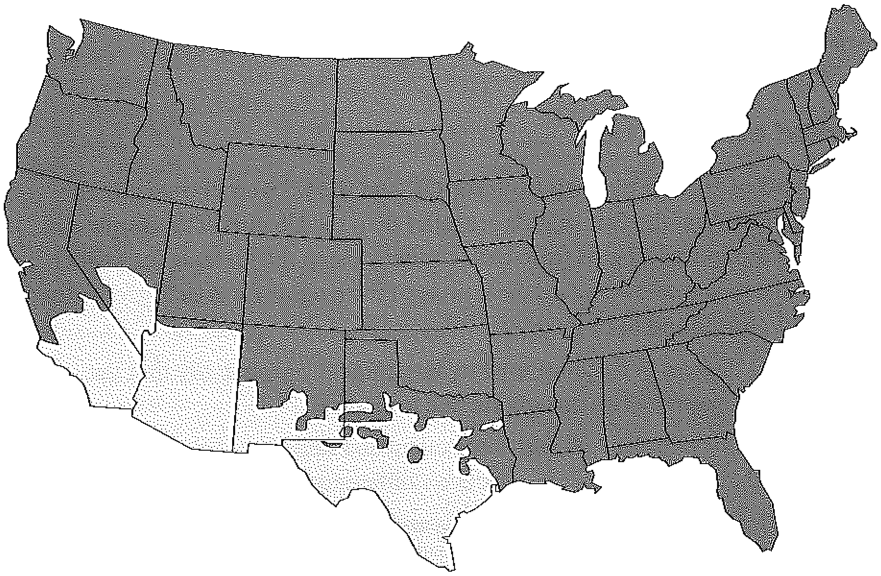


Figure 1.5 AHBs have not moved beyond the areas shaded yellow in the last 10 years. So, states in the south east (Louisiana, Florida, Alabama and Mississippi) seem unlikely to be bothered by AHBs if the 137.5cm (55 inches) of rain correlation holds true. This is an example of a mathematical correlation and is not evidence of a cause. In order to find out if this is a cause, scientists must design experiments to explain mechanisms which may be the cause of the observed correlation.

We can see that a correlation exists but we do not know why it exists. We could do some experiments to discover the cause.

- Is it due to the fact that too much rain will flood the nests of the bees?
- Does too much rain cause a fungus to grow on the bee hives which destroys any colony of bees living in areas with more than an annual rainfall of more than 137.5 cm (55 inches)?

Scientific experiment must be done to show the cause of the correlation that we are measuring.

Cormorants

When using a mathematical correlation test, the value of *r* signifies the correlation. The value of *r* can vary from +1 (completely positive correlation) to 0 (no correlation) to –1 (completely negative correlation). For an example, we can measure in millimetres the size of breeding cormorant birds to see if there is a correlation between the sizes of males and females which breed together.

Pair numbers	Size of female cormorants	Size of male cormorants
1	17.1	16.5
2	18.5	17.4
3	19.7	17.3
4	16.2	16.8
5	21.3	19.5
6	19.6	18.3
$r = 0.88$		

The *r* value of 0.88 shows a positive correlation between the sizes of the two sexes: large females mate with large males. However, correlation is not cause. To find the cause of this observed correlation requires experimental evidence. There may be a high correlation but only carefully designed experiments can separate causation from correlation.

Some questions

- For years we have known that there is a high positive correlation between smoking and lung cancer. Does this high positive correlation prove that smoking causes lung cancer?

  No, it does not prove that smoking causes lung cancer. Only data collected from a well designed experiment can show cause.

- How can the cause of lung cancer be determined?

  For many years scientists have performed experiments to study the cause of the correlation which was originally observed. Currently, scientific evidence shows that smoking increases the chances of contracting lung cancer. Experiments show cause.

- Can you guess what might be the cause of the positive correlation in the mating of large female cormorants with large male cormorants?

  It is possible that the larger females scare off the smaller males. We can only know the answer to this by designing an experiment which studies the behaviour of the birds during mating.