SMARTER STUDY SKILLS

HOW TO USE STATISTICS

STEVE LAKIN

ALWAYS LEARNING

PEARSON

HOW TO USE STATISTICS

The mean in sigma notation

Recall that the mean of a list of data is worked by adding up all the values, and then dividing by the number of values.

Suppose we have a list of data x with terms $x_1, x_2, x_3, \ldots, x_n$, and so there are n terms (n pieces of data). Then we want to add up all these x_i and then divide by n (the number of terms) and so we can give the formula for the mean in sigma notation as follows:

> The mean of the list of data $x = x_1, x_2, x_3, \ldots, x_n$ is $\dfrac{\sum_{i=1}^{n} x_i}{n}$, or it can be written as $\dfrac{1}{n}\sum_{i=1}^{n} x_i$, which is perhaps slightly easier to write without using up too much vertical space.

Look at this carefully and convince yourself that all it is saying is 'add up the values and then divide by the number of values', which is the definition of the mean. We are going to use sigma notation a lot, so make sure you understand it and how it works, so that you can be confident looking at formulae that use it.

> It is common to denote the mean of a list of data x as $\overline{x}$: that is, a bar above the x. Remember this notation, as we shall use it again.

The variance in sigma notation

Note: In everything we do here, we are considering the variance and standard deviation to be the population variance and population standard deviation. The sample variance and sample standard deviation are almost exactly the same, but with $n - 1$ instead of n.

Recall that the variance is worked out by taking the difference between every value and the mean, and squaring them, and then adding them all up and dividing by the number of values. Using the notation $\overline{x}$ for the mean as discussed above, we can create a formula for the variance.

The variance of the list of data $x = x_1, x_2, x_3, \ldots, x_n$ is $\dfrac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n}$, or it can be written as $\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$, which is again perhaps slightly easier to write without using up too much vertical space.

Look carefully at this formula and try to realise what it is saying. For each value, take the difference from the mean, and square the answer. Then add up all the values, and divide the total by n. This is exactly the definition of the variance, but formalised in mathematical notation.

An alternative formula for the variance

It is possible to show that this formula for the variance is the same as the formula $\dfrac{\sum_{i=1}^{n}x_i^2}{n} - \overline{x}^2$ (or $(\frac{1}{n}\sum_{i=1}^{n}x_i^2) - \overline{x}^2$), which gives us a faster way to work out the variance: we add up all the squares of the values and divide by n, and then subtract the mean squared. The advantage of this formulation is that you don't have to take the mean away at every step; just take away its square at the end.

I won't prove here why these are equivalent; but if you are confident in your mathematical ability, then feel free to have a go. But I will illustrate by example:

Example

Work out the variance of the set of data 4, 5, 9, 10.

Solution 1

Using the first formulation of the variance $\dfrac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n}$, we need to take the difference of every value from the mean, and then divide by n, the number of values, which is 4 in this case. The mean is $\dfrac{4+5+9+10}{4} = \dfrac{28}{4} = 7$. Hence we need to work out the squared differences and divide by 4 (the number of values), and so work out $\dfrac{(4-7)^2 + (5-7)^2 + (9-7)^2 + (10-7)^2}{4} = \dfrac{(-3)^2 + (-2)^2 + 2^2 + 3^2}{4} =$

$$\frac{9 + 4 + 4 + 9}{4} = \frac{26}{4}$$

which is 6.5 as a decimal, and so the variance is 6.5.

Solution 2

The second formulation of the variance is $\frac{\sum_{i=1}^{n} x_i^2}{n} - \bar{x}^2$, so we add up all the squares of the values, divide by n, and then take away the mean squared. So, in this example (remember that the mean is 7), we work out $\frac{4^2 + 5^2 + 9^2 + 10^2}{4} - 7^2$, which is $\frac{16 + 25 + 81 + 100}{4} - 49$

$$= \frac{222}{4} - 49 = 55.5 - 49 = 6.5$$

and we get the same answer for the variance.

You can use either formulation. The second is probably a little easier to calculate, as you don't have to take the mean away at every step, but the first is probably more 'natural' in terms of what variance *means* (the distance of values away from the average).

The standard deviation in sigma notation

Remembering that the standard deviation is simply the square root of the variance, then the formula for the standard deviation is just

$$\sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}} \quad \text{or} \quad \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n} - \bar{x}^2}$$

depending on which version of the variance formula you prefer to use.

Pi notation

There is a similar notation to sigma notation that is used when you multiply things together, rather than add them up. This notation uses a capital 'pi' symbol (again, a Greek letter), and so for example $\prod_{i=1}^{5} i$ means $1 \times 2 \times 3 \times 4 \times 5$ (which is 120). You probably won't encounter this again (it's nothing like as common as sigma notation), but I'm mentioning just in case you do come across it and need to know what it means.

Summary

Sigma notation takes some getting used to, but it is a very concise, effective way of representing series, which crop up all the time in statistics. Practise with it and become comfortable with what something in sigma notation means, and you will see that you can write quite complex formulae in a concise mathematical way, which is very important when it comes to bigger problems.

Exercises

1 Write down the next term in each of the following sequences:

Example: 2, 5, 8, 11, 14, ...

Solution: In this sequence, you can spot that the difference between each term is 3, so the next term is 17.

(a) 1, 5, 9, 13, 17, ...

(b) 1, 4, 9, 16, 25, ...

(c) 1, 10, 100, 1000, ...

(d) $\frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \ldots$

(e) 1, 3, 9, 27, ...

(f) 1, 1, 2, 3, 5, 8, 13, 21, ...

(*Hint for (f): this is known as the Fibonacci sequence, so research this if you can't spot the pattern.*)

2 Write out the following series (given in sigma notation) in full. Can you actually perform the addition and give an answer?

Example: $\sum_{i=1}^{5} 3i$

Solution: First i is 1, so we write down the value of the expression $3i$, which is $3 \times 1 = 3$. Then we consider $i = 2$, so we write down 3×2 (since $i = 2$), which is 6. Similarly, when $i = 3$, we write down 9; when $i = 4$, we write down 12; and when $i = 5$ (the stopping point) we write down 15. Hence this series is $3 + 6 + 9 + 12 + 15$, which you can calculate to be 45.

(a) $\sum_{i=1}^{6} 4i$

(b) $\sum_{i=1}^{4} (2i + 1)$

(c) $\sum_{i=0}^{4} 2^i$

(d) $\sum_{i=5}^{10} (i + 1)^2$

(e) $\sum_{i=0}^{\infty} \frac{1}{2^{i+1}}$

(f) $\sum_{i=1}^{\infty} i$

3 Use both formulae for the variance to calculate the variance and standard deviation of the following lists of data.

Example: 1, 3, 6, 7, 8

Solution: The mean is $\frac{1+3+6+7+8}{5} = \frac{25}{5} = 5$. Using the first formula, we work out the difference between every value and the mean and square them, before dividing by the number of values (5), so we work out the variance to be

$$\frac{(1-5)^2+(3-5)^2+(6-5)^2+(7-5)^2+(8-5)^2}{5}$$
$$= \frac{16+4+1+4+9}{5} = \frac{34}{5} = 6.8.$$

Using the second formula, we square all of the values, divide by the number of values (5), and subtract the mean squared, and so we work out the variance to be

$$\frac{1^2+3^2+6^2+7^2+8^2}{5} - 5^2 = \frac{1+9+36+49+64}{5} - 25 =$$
$$\frac{159}{5} - 25 = 31.8 - 25 = 6.8.$$

Either way, we work out the variance to be 6.8, and so the standard deviation is $\sqrt{6.8} = 2.608$ (to 3 decimal places)

(a) 1, 4, 7, 9, 10, 13

(b) 2, 3, 8, 9, 10

(c) 2, 5, 5, 5, 6, 8, 9, 10

(d) 4, 4, 4, 4, 4, 4

4 Write the following series using sigma notation:

Example: $3 + 6 + 9 + 12 + 15$.

Solution: The ith term is $3i$ (the first term is 3×1, the second term is 3×2, etc.) and so, since there are five terms, this is $\sum_{i=1}^{5} 3i$.

(a) $4 + 8 + 12 + 16 + 20 + 24 + 28$

(b) $1 + 2 + 3 + 4 + 5 + \ldots$

(c) $1 + 3 + 5 + 7 + 9 + 11$

(d) $2 + 1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \ldots$