Lecture 7

Correlation example

Data:

i	1	2	3	4	5	6
x	1	2	3	4	10	10
y	1	3	3	5	1	11

Formula for calculating correlation

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

Alternative Formula

$$r = \frac{1}{n-1} \frac{1}{s_x} \frac{1}{s_y} \left(\sum_{i=1}^{n} x_i y_i - n \bar{x} \bar{y} \right)$$

Where

- r – correlation
- n – number of observations
- $\bar{x}$ – sample mean of x values
- $\bar{y}$ – sample mean of y values
- s_x – sample standard deviation of x values
- s_y – sample standard deviation of y values

Calculate the components of correlation formula

$n = 6$

Sample means

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{1+2+3+4+10+10}{6} = \frac{30}{6} = 5$$

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{1+3+3+5+1+11}{6} = \frac{24}{6} = 4$$

Next we need the standard deviations

$$s_x = \sqrt{\frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{n-1}}$$ ← formula for standard deviation

$$s_y = \sqrt{\frac{\sum_{i=1}^{n} y_i^2 - n\bar{y}^2}{n-1}}$$

Already have $\bar{y}, \bar{x}, n$ from above. Need $\sum x^2, \sum y^2$

$$\sum_{i=1}^{n} x_i^2 = 1^2 + 2^2 + 3^2 + 4^2 + 10^2 + 10^2$$
$$= 1 + 4 + 9 + 16 + 100 + 100$$
$$= 230$$

$$\sum_{i=1}^{n} y_i^2 = 1^2 + 3^2 + 3^2 + 5^2 + 1^2 + 11^2$$
$$= 1 + 9 + 9 + 25 + 1 + 121$$
$$= 166$$

So

$$s_x = \sqrt{\frac{230 - 6(5)^2}{6-1}} = \sqrt{\frac{230-150}{5}} = \sqrt{16} = 4$$

$$s_y = \sqrt{\frac{166 - 6(4)^2}{6-1}} = \sqrt{\frac{166-96}{5}} = \sqrt{14} = 3.7416 \text{ (4dp)}$$

Lastly we need $\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$ (or $\Sigma x_i y_i - n\bar{x}\bar{y}$)

$$\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$$
$$= (1-5)(1-4)$$
$$+(2-5)(3-4)$$
$$+(3-5)(3-4)$$
$$+(4-5)(5-4)$$
$$+(10-5)(1-4)$$
$$+(10-5)(11-4)$$
$$= (-4)(-3)+(-3)(-1)$$
$$+(-2)(-1)+(-1)(1)$$
$$+(5)(-3)+(5)(7)$$
$$= 12+3+2-1-15+35$$
$$= \underline{36}$$

$$\Sigma x_i y_i = 1\times1 + 2\times3 + 3\times3 + 4\times5 + 10\times1 + 10\times11$$
$$= 1 + 6 + 9 + 20 + 10 + 110$$
$$= 156$$

$$\Sigma x_i y_i - n\bar{x}\bar{y}$$
$$= 156 - 6(5)(4)$$
$$= 156 - 120$$
$$= \underline{36}$$

this method is quicker for hand calculation

Check that these agree

Now substitute back into correlation formula

$$r = \frac{1}{n-1} * \frac{1}{s_x} * \frac{1}{s_y} * \left(\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} \right)$$

$$= \frac{1}{6-1} * \frac{1}{4} * \frac{1}{3.7416} * 36$$

$$= 0.48 \quad (2dp)$$

We would say that x and y are weakly positively correlated.

Probability

Probability is based on the idea of a random experiment. That is an experiment for which we cannot predict the outcome before it is carried out.

A phenomenom is random if the outcome can not be predicted but it has a regular distribution in very many repetitions of the experiment.

The probability of an event is the proportion of times that it occurs in very many repetitions of a random experiment.

In class experiment

We will carry out a random experiment using thumb tacks

Experimental procedure:

1. Drop a thumb tack from about 6 inches above a flat surface
2. Record whether it lands up or down 
3. Repeat many times to estimate the probability of the thumb tack landing up.

How to record results

Use a table. Suppose we get the following sequence of tosses: U, U, D, U, D, U, ...

then ⇒ Toss #	Cumulative # UP	Current estimate of P(UP)
1	1	1/1 = 1
2	2	2/2 = 1
3	2	2/3 = 0.66
4	3	3/4 = 0.75
5	3	3/5 = 0.6
6	4	4/6 = 0.66
⋮ (many trials)	⋮	⋮
n	n_{up}	$\frac{n_{up}}{n} = P(UP)$.