1

STAT2008/STAT6038

Revision

What is a population? What is a sample?

2

- **Population:** a collection of the whole of something – e.g. all people who live in Belconnen
- **Sample:** a set of individuals drawn from a population e.g. the people who live in Macgregor are a sample of all people who live in Belconnen.

If we have a population….

3

- We can get parameters – true values for things like the centre and spread of the population
- We know the answers – what proportion are this tall? We look at the population and get the answer.

If we have a sample…

4

- We can get statistics – these are values that estimate the parameters e.g. sample centre and sample spread used to estimate population centre and population spread

Types of Data

5

- Two basic types of data – discrete or continuous
- Discrete data – nominal, ordinal, count
- Continuous data – also called interval
- Other sorts of information – e.g. comments in interview/survey – qualitative

Discrete data examples

6

- Nominal – faculty of study, eye colour, job
- Ordinal – rank teaching as poor/fair/good/very good
- Count data – number of people at a party

Continuous data examples

7

- Anything *measured*
  - Height
  - Weight
  - Exam marks
  - Incomes
  - Prices
- "to the nearest ______"

Sample Mean

8

- Arithmetic mean – average
- If observations are labelled $X_1, X_2, \ldots, X_n$ then sample average is called $\overline{X}$
- Calculated as

$$\overline{X} = \frac{1}{n}\left(X_1 + X_2 + \ldots + X_n\right)$$
$$= \frac{1}{n}\sum_{i=1}^{n} X_i$$

Easy example - mean

9

- Data: 5, 7, 1, 2, 4

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$
$$= \frac{1}{5}\left(5+7+1+2+4\right)$$
$$= \frac{1}{5} * 19$$
$$= 3.8$$

Measures of variability

10

- Need to know how "spread out" the data are
- Range (maximum obs-minimum obs)
- IQR (Q_3-Q_1)
- Variance/Standard Deviation
- Coefficient of variation
- **Note:** IQR much less influenced by extreme values than variance/std dev/cv

Variance

11

- A measure of spread of data
- Measured in ***square units***
- Sample value is given symbol s^2
- Calculated as

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$
$$= \frac{1}{n-1}\left[\sum_{i=1}^{n} X_i^2 - n\overline{X}^2\right]$$

Easy example - variance

12

- Data: 5, 7, 1, 2, 4

$$s^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} X_i^2 - n\overline{X}^2\right]$$
$$= \frac{1}{4}\left[\left(25+49+1+4+16\right) - 5*3.8^2\right]$$
$$= \frac{1}{4}\left[\left(95\right) - 5*14.44\right]$$
$$= \frac{1}{4}*22.8 = 5.7$$

Standard Deviation

13

- Another measure of spread
- Given symbol s
- Calculated as $s = \sqrt{s^2}$
- Standard Deviation is the square-root of variance

Easy example – standard deviation

14

- Data: 5, 7, 1, 2, 4

$$s^2 = 5.7$$

$$s = \sqrt{s^2} = \sqrt{5.7} = 2.387 \text{ (to 3dp)}$$

- Measured in same units as original data.

Populations vs Samples

15

- Population – every individual of a certain type
- Sample – selection of individuals
- Mostly, we are dealing with samples
- Populations have parameters – certain true values which describe them. E.g. if we measure every individual, we can calculate the exact average and exact variance of the population. These are called population parameters.
- From a sample we calculate statistics, or estimates of the parameters. E.g. from a sample we can estimate the true population mean by using the same mean; if we have a sample standard deviation, we can estimate the population standard deviation.

Sample vs population

16

	Sample (Real World)	Population (Fantasy)
Average (Mean)	$\bar{X}$	μ
Variance	s^2	σ^2
Standard Deviation	s	σ

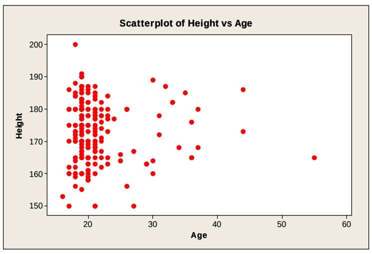
If we have two measurements on one observation…

17

- E.g. height and weight of a person, weekly income and amount spent on rent per week.
- Scatterplot
- Covariance
- Correlation

Scatterplot – Always plot your data!!!

18



Covariance

19

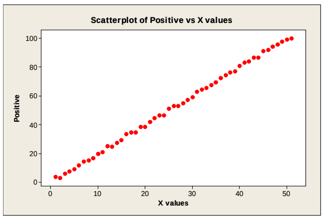
- Measures the linear relationship between X and Y – sign indicates direction of slope, but magnitude is dependent on units of measurement (so cannot indicate strength of relationship).
- Calculated as

$$\operatorname{cov}(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$
$$= \frac{1}{n-1}\left[\sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}\right]$$

Values of covariance

20

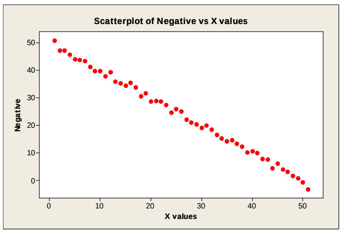
- If cov>0, then as X increases, Y increases; as X decreases, Y decreases (positive slope)



Values of covariance

21

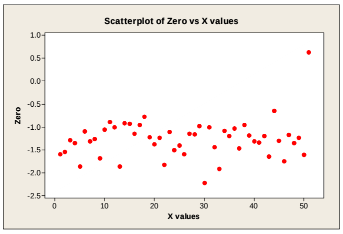
- If cov<0, then as X increases, Y decreases; as X decreases, Y increases (negative slope)



Values of covariance

22

- If cov=0, then as X changes, Y doesn't change ➔ variables are not linearly related



Coefficient of Correlation

23

- Also measures strength of linear relationship between X and Y.
- Is bounded between -1 and +1.
- Calculated as

$$\rho = \frac{COV(X,Y)}{\sigma_X \sigma_Y}, \qquad r = \frac{\operatorname{cov}(X,Y)}{s_X s_Y}$$

If correlation equals….

24

- If r=-1, perfect negative linear relationship
- If r=+1, perfect positive linear relationship
- If r=0, no LINEAR relationship