Z Scores & Correlation

Greg C Elvers

1

Z Scores

- A *z score* is a way of standardizing the scale of two distributions
- When the scales have been standardize, it is easier to compare scores on one distribution to scores on the other distribution

2

An Example

- You scored 80 on exam 1 and 75 on exam 2. On which exam did you do better?
- The answer may *or may not* be that you did better on exam 2
- In order to decide on which exam you did better, you must also know the mean and standard deviation of the exams

3

An Example

- The mean and standard deviation of Exam 1 were 85 and 5, respectively
- The mean and standard deviation of Exam 2 were 70 and 5, respectively
- So, you scored below the mean on exam 1 and above the mean on exam 2
- On which exam did you do better?

4

Z Scores

- A z score is defined as the deviate score (the observed score minus the mean) divided by the standard deviation
- It tells us how far a score is from the mean in units of the standard deviation

$$z = \frac{\left(X - \overline{X}\right)}{s}$$

5

An Example

- You have a z score of -1 on the first exam
  - Your score was one standard deviation below the mean on exam 1
- You have a z score of 1 on the second exam
  - Your score was one standard deviation above the mean on exam 2
- You did better on exam 2

$$z = \frac{\left(X - \overline{X}\right)}{s} = \frac{(80 - 85)}{5} = -1$$

$$z = \frac{\left(X - \overline{X}\right)}{s} = \frac{(75 - 70)}{5} = 1$$

6

Important Properties of Z Scores

- The mean of a distribution of z scores is always 0
- The standard deviation of a distribution of z scores is always 1
- The sum of the squared z scores always equals N

$$\mu_Z = 0$$

$$\sigma_z = 1$$

$$\sum z^2 = N$$

7

Proofs

$$\mu_z = 0$$

$$\frac{\sum\left(\frac{X-\mu}{\sigma}\right)}{N} = 0$$

$$\frac{\frac{1}{\sigma}\sum(X-\mu)}{N} = 0$$

$$\frac{\frac{1}{\sigma}\left(\sum X - \sum \mu\right)}{N} = 0$$

$$\frac{\frac{1}{\sigma}\left(\sum X - N\cdot\mu\right)}{N} = 0$$

$$\frac{\frac{1}{\sigma}\left(\sum X - N\frac{\sum X}{N}\right)}{N} = 0$$

$$\frac{\frac{1}{\sigma}\left(\sum X - \sum X\right)}{N} = 0$$

$$\frac{\frac{1}{\sigma}0}{N} = 0$$

$$0 = 0$$

$$\sum z^2 = N$$

$$\sum\left(\frac{X-\mu}{s}\right)^2 = N$$

$$\sum\frac{(X-\mu)^2}{s^2} = N$$

$$\frac{1}{s^2}\sum(X-\mu)^2 = N$$

$$\frac{1}{\frac{\sum(X-\mu)^2}{N}}\sum(X-\mu)^2 = N$$

$$\frac{N}{\sum(X-\mu)^2}\sum(X-\mu)^2 = N$$

$$N = N$$

$$\sigma_z = 1$$

$$\sigma_z^2 = 1$$

$$\sigma_z^2 = \frac{\sum(z-\mu_z)^2}{N} = 1$$

$$\mu_z = 0$$

$$\sigma_z^2 = \frac{\sum z^2}{N} = 1$$

$$\sum z^2 = N$$

$$\sigma_z^2 = \frac{N}{N} = 1$$

8

Z scores and Pearson's r

- Pearson's r is defined as:

$$r = \frac{\sum z_x z_y}{N}$$

What the Formula Means

- The z scores in the formula simply standardize the unit of measure in both distributions
- The product of the z scores is maximized when the largest z_x is paired with the largest z_y

r = 1

- Because of the unit standardization, when there is a perfect correlation $z_x = z_y$
- Then $z_x z_y = z_x^2 = z_y^2$

$$r = \frac{\sum z_x^2}{N} = \frac{N}{N} = 1$$

r = 0

- When r = 0, large z_x can be paired with large or small z_y
- Furthermore, positive z_x can be paired with either positive or negative z_y
- The sum of $z_x z_y$ will tend to 0
- Thus, r will tend to 0

Computational Formula for r

$$r = \frac{\sum XY - \frac{\left(\sum X\right)\left(\sum Y\right)}{N}}{\sqrt{\left[\sum X^2 - \frac{\left(\sum X\right)^2}{N}\right]\left[\sum Y^2 - \frac{\left(\sum Y\right)^2}{N}\right]}}$$

Coefficient of Determination

- The *coefficient of determination* is the proportion of variance in one variable that is explainable by variation in the other variable
- It tells us how well we can predict the value of one variable given the value of another

Coefficient of Determination

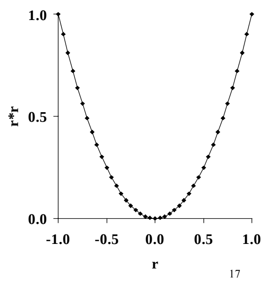
- When there is a perfect correlation between two variables, then all the variation in one variable can be explained by variation in the other variable
- Thus the coefficient of determination must equal 1

Coefficient of Determination

- When there is no relation between two variables, then none of the variation in one variable can be explained by variation in the other variable
- Thus the coefficient of determination must equal 0

Coefficient of Determination

- The coefficient of determination is defined as r^2
- When r = 1 or r = -1, $r^2 = 1$, as it should be
- When r = 0, $r^2 = 0$, as it should be



Coefficient of Nondetermination

- The *coefficient of nondetermination* is the amount of variation in one variable that is *not* explainable by the variation in the other variable
- The coefficient of nondetermination equals $(1 - r^2)$

Correlation and Causation

- Correlation *does not show causation*
- Just because two variables are correlated (even perfectly correlated) does not imply that changes in one variable cause the changes in the other variable
- E.g., even if drinking and GPA are correlated, we do not know if people drink more because their GPA is low (drink to alleviate stress) or if drinking causes one's GPA to be low (less study time) or neither of these

Correlation and Causation

- There is always a chance that the variation in both variables is due to the variation in some third variable
- r = 0.95 for number of storks sighted in Oldenburg Germany and the population of Oldenburg from 1930 to 1936
  - Storks do not cause babies
  - Babies do not cause storks
  - What is the third variable that causes both?

Special Correlation Coefficients

Scale	Symbol	Used With
Nominal	r_{phi} (phi coefficient)	2 dichotomous variables
	r_b (biserial r)	1 dichotomous variable with underlying continuity; one variable can take on more than 2 values
	r_t (tetrachoric)	2 dichotomous variables with underlying continuity
Ordinal	r_s (Spearman r)	Ranked data (both variables at least ordinal)
	τ (Kendall's tau)	Ranked data
Interval or Ratio	Pearson r	Both variables interval or ratio
	Multiple r	More than 2 interval or ratio scaled variables