July 21 - 24, 2021 | Garki Hospital Abuja | Resource Person

Research Methodology Boot Camp

with Epi Info Training

Dr. Adamu Onu
MBBS, FWACP (FM)
MS Epidemiology & Biostatistics
PhD Public Health (Epidemiology)

Target Audience Clinical Researchers, Post-Part 1 Residents, and Others

Important Information

- Limited slots are available on a first come, first served basis
- Laptop running Windows 10 required
- Organized as morning lecture sessions and afternoon hands on coaching sessions

Highlights

- Research Methodology
- Research Design
- Data Management
- Sample Size Calculations
- Test Statistics
- Interpretation of Results
- Report Writing
- Hands-on training sessions
- Statistical consulting sessions

For further details contact
Email: epimetrix@gmail.com
Phone: +234 803 474 9930

epimetrix
health systems
data for health decisions

Correlation

Correlation

- **Correlation analysis** can be calculated to investigate the **linear relationship** of variables.
- How strong the **correlation** is is determined by the **correlation coefficient**, which varies from –1 to +1.
- Correlation analyses can thus be used to make a statement about the strength and direction of the correlation.

> You want to find out if there is a connection between height and weight

- If the correlation analysis shows that two characteristics are related, one characteristic can be used to predict the other.
- If the correlation mentioned in the example is confirmed, it is then possible to predict weight by the height using a linear regression.

Correlation and causality

> If the correlation between height and weight is analysed and a strong correlation occurs, it would be logical to assume that weight is influenced by the height (and not vice versa), but this assumption can by no means be proven on the basis of a correlation analysis.

- Correlations need not be **causal relationships**.
- Any correlations should be investigated more closely, but never interpreted immediately.
- It can happen that the correlation between variable *x* (height) and *y* (weight) is generated by the variable *z* (caloric intake) — *Partial Correlation*

Interpret correlation

- With the help of correlation analysis two statements can be made, one about
  - the direction and
  - the strength
- of the linear relationship between two scale or ordinal variables.
- The direction indicates whether there is a **positive** correlation or a **negative** correlation.

Positive correlation

- A positive correlation exists if larger values of variable A are accompanied by larger values of variable B.
- Height and weight, for example, correlate positively and a correlation coefficient of between 0 and 1 results, i.e. a positive value.

Negative correlation

- A negative correlation exists if larger values of variable A are accompanied by smaller values of variable B.
- Product price and the sales quantity usually have a negative correlation: the more expensive a product is, the smaller the sales quantity.
- In this case, the correlation coefficient is between –1 and 0, so it assumes a negative value.

Test correlation for significance

- The significance of correlation coefficients can be tested using a *t* test.
- As a rule, it is tested whether the correlation coefficient is significantly different from zero, i.e. linear independence is tested.
- In this case, the null hypothesis is that there is no correlation between the variables under consideration.
- In contrast, the alternative hypothesis assumes that there is a correlation.

Test correlation for significance

$$t = \frac{r \cdot \sqrt{n-2}}{\sqrt{1-r^2}}$$

- Where *n* is the sample size and *r* is the determined correlation in the sample.
- The corresponding *p* value can be easily calculated using a *t* distribution with *n* – 2 degrees of freedom.

Pearson correlation

- With the Pearson correlation analysis you get a statement about the linear correlation between scale variables.
- The respective covariance is used for the calculation.
- The covariance gives a positive value if there is a positive correlation between the variables and a negative value if there is a negative correlation.

Pearson correlation

- The covariance is calculated using :

$$Cov(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{N-1}$$

- However, the covariance is not normalized and can assume values between $-\infty$ and $+\infty$ ($-\infty \leq Cov(x,y) \leq \infty$).
- This makes it difficult to compare the strength of relationships between different variables.

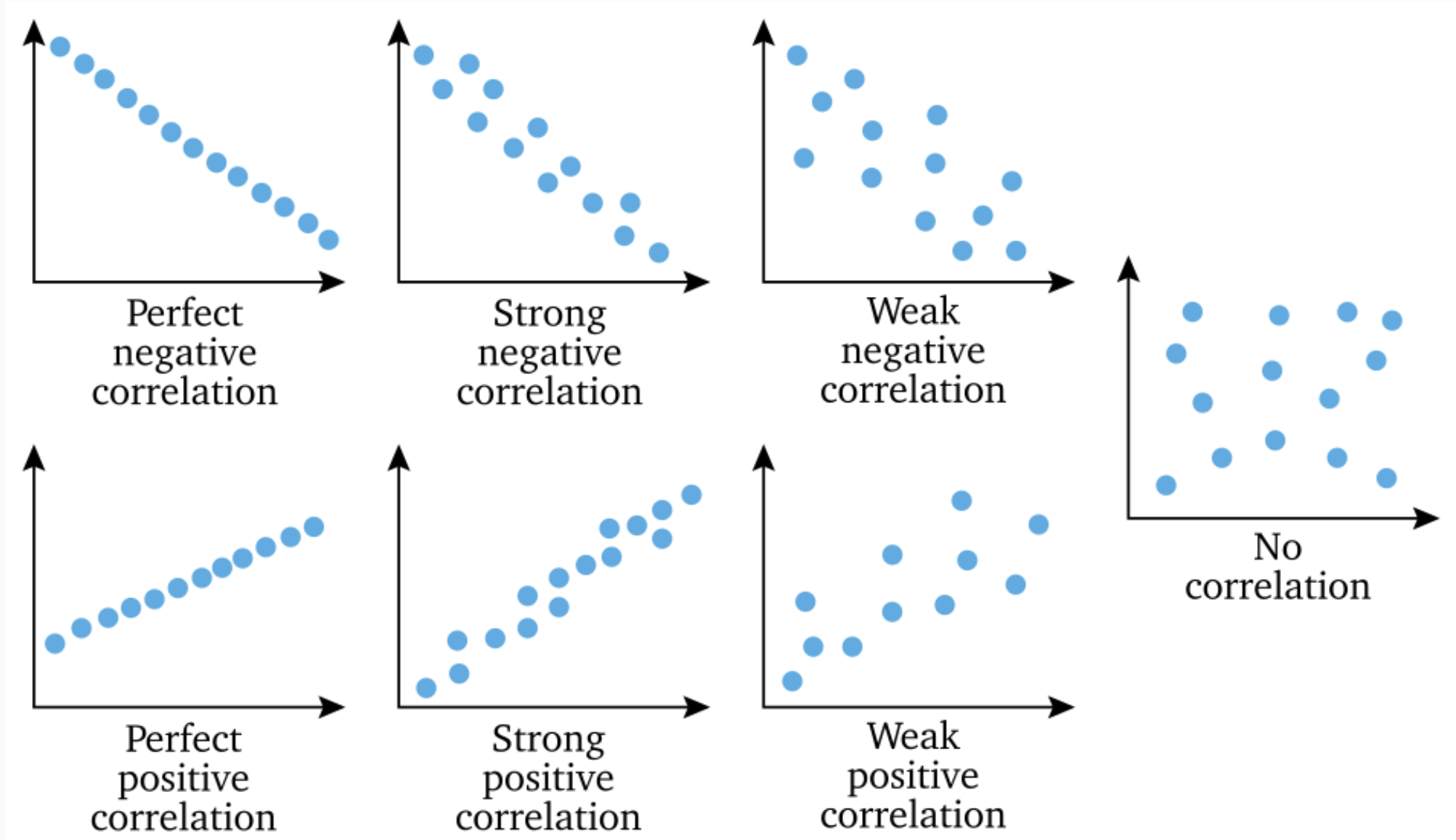
Pearson correlation

- For this reason, the **correlation coefficient**, also called **product-moment correlation**, is calculated
- The correlation coefficient is obtained by normalizing the covariance.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Pearson correlation

- The Pearson correlation coefficient (r) can now take values between -1 and $+1$ ($-1 \leq r \leq +1$):
  - The value $+1$ means that there is an entirely positive linear relationship (the more, the more).
  - The value -1 indicates that an entirely negative linear relationship exists (the more, the less).
  - With a value of 0 there is no linear relationship, i.e. the variables do not correlate with each other.


Perfect negative correlation
Strong negative correlation
Weak negative correlation
No correlation
Perfect positive correlation
Strong positive correlation
Weak positive correlation

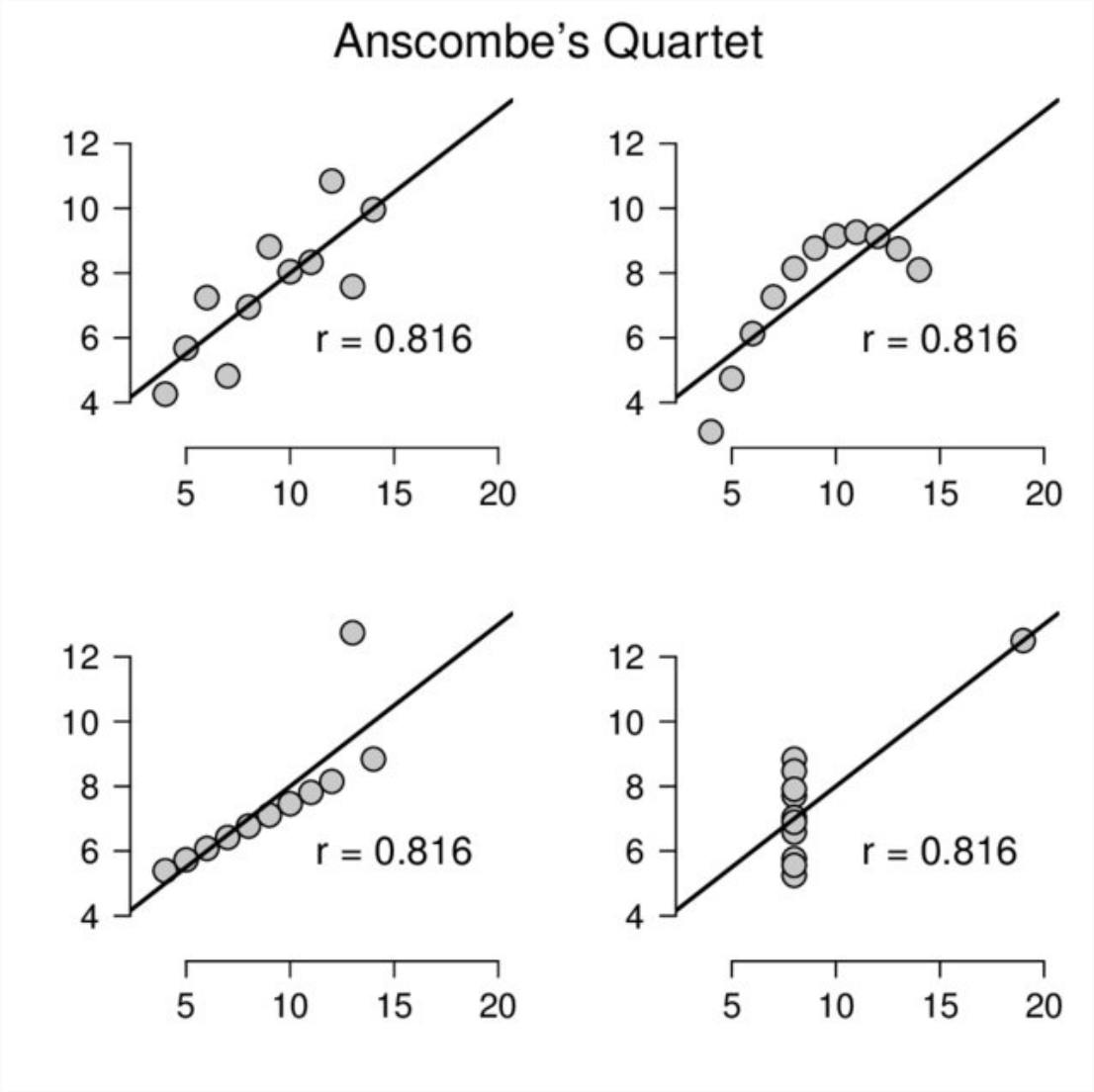
Strength of correlation

- With regard to the strength of the correlation, the following table can be taken as a guide:

Amount of r	Strength of correlation
0.0 < 0.1	no correlation
0.1 < 0.3	little correlation
0.3 < 0.5	medium correlation
0.5 < 0.7	high correlation
0.7 < 1.0	very high correlation

Use scatter plots to visually check in advance whether a linear relationship exists.

The Pearson correlation is only useful and purposeful if linear relationships are present.



Assumptions

- Both x and y must be normally distributed
- There must be a linear relationship between the variables

Spearman rank correlation

- Spearman correlation analysis is used to calculate the relationship between two variables that have **ordinal** level of measurement.
- Spearman rank correlation is the non-parametric equivalent of Pearson correlation analysis.
- This procedure is therefore used when the prerequisites for a Pearson correlation analysis are not met.

Spearman rank correlation

- The calculation of the rank correlation is based on **ranking** the data.
- Measured values are not used for the calculation, but are transformed into ranks.
- The test is then performed using these ranks.
- For the rank correlation coefficient (ρ), values between –1 and 1 are possible.
  - If $\rho < 0$, there is a negative linear correlation.
  - If $\rho > 0$, there is a positive linear relationship.
  - If $\rho = 0$, there is no relationship.

Height	Weight	Age
1.62	53	20
1.72	71	30
1.85	85	25
1.82	86	24
1.72	76	23
1.55	62	25
1.65	68	26
1.77	77	20
1.83	97	33
1.53	65	24

A student wants to know if there is a correlation between height, weight and age among the participants in the statistics course.

	Height	Weight	Age
Height	1.00	0.86	0.28
Weight	0.86	1.00	0.52
Age	0.28	0.52	1.00

Partial correlation

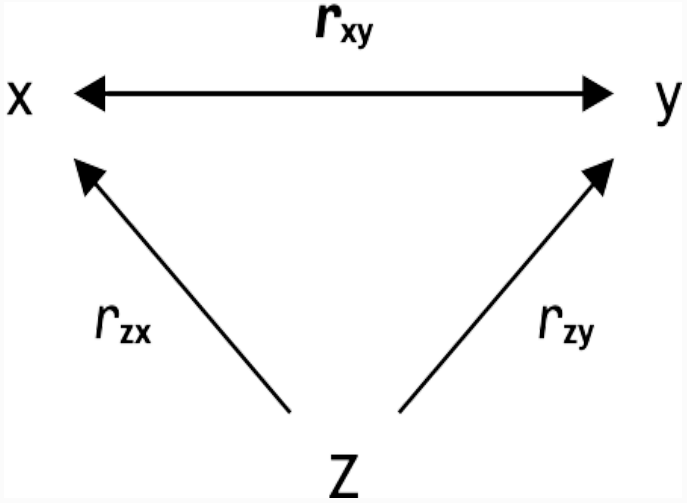
- **Partial correlation** calculates the correlation between two variables to the exclusion of a third variable.
- This makes it possible to find out whether the correlation r_{xy} between variables *x* and *y* is generated by the variable *z*.
- Is the correlation between variable *x* (height) and *y* (weight) generated by the variable *z* (caloric intake)?

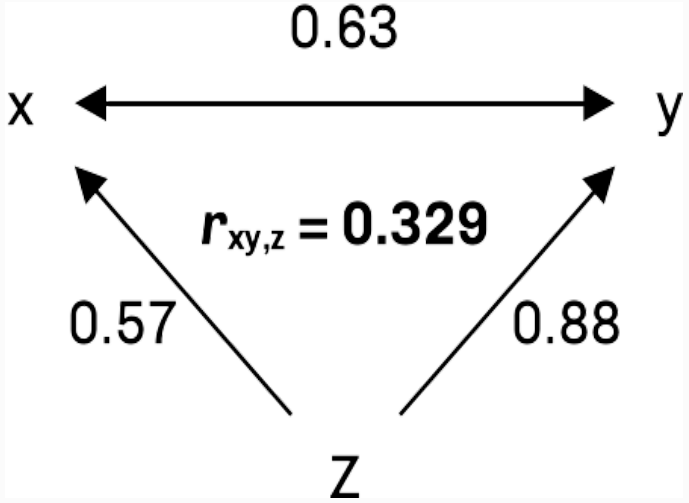
Calculation of partial correlation

- The partial correlation $r_{xy,z}$ tells how strongly the variable x correlates with the variable y, if the correlation of both variables with the variable z is calculated out.

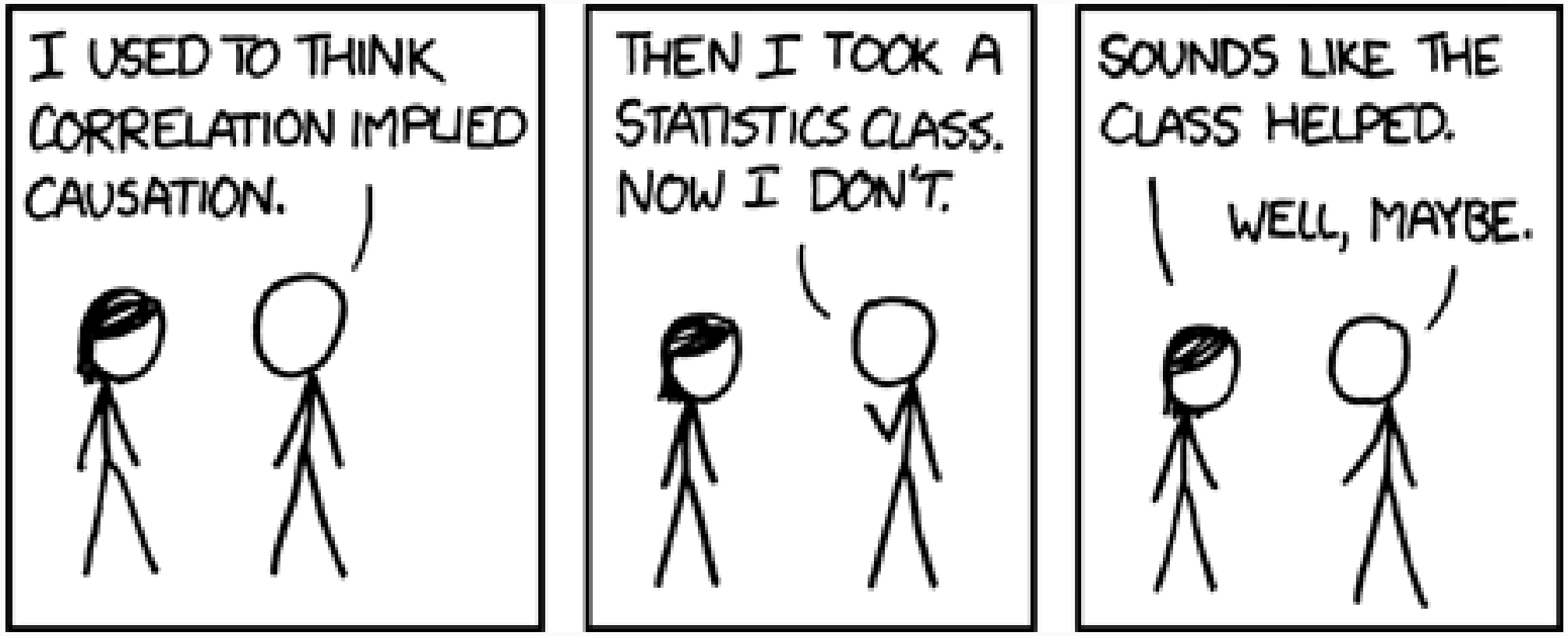
$$r_{xy,z} = \frac{r_{xy} - r_{xz} \cdot r_{yz}}{\sqrt{(1 - r_{xz}^2) \cdot (1 - r_{yz}^2)}}$$

- r_{xy} = Correlation of x with y
- r_{xz} = Correlation of z with x
- r_{yz} = Correlation of z with y





- The correlation between *x* and *y* is only 0.329 when the correlations with *z* is excluded.


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.