

Error in Epidemiological Studies

Interpretation of an Association

An observed association between exposure and outcome can be the result of:

- Bias (systematic error)
- Confounding/Effect Modification
- Chance (random error)
- Causal

Interpretation of an Association

- The observed statistical association between a certain outcome and the hypothesized exposure could be a matter of chance.
- Or it could be the result of systematic errors in collection of data (sampling, disease and exposure ascertainment)
- or its interpretation: the role of bias Or it could be due to the effect of additional variables that might be responsible for the observed association: the role of confounding Or it could be a real association

Dealing with error

To achieve the research goals we consider:

Precision

Validity

Precision is affected by random error and is dealt with by statistical procedure. We deal with this in Sampling and in Hypothesis Testing.

Validity is elimination of systematic error (bias). Systematic error includes issues of:

Selection bias

Information bias

Validity

- *Internal Validity*
 - How well was study done?
 - Are measurements valid and accurate?
 - Are groups comparable?
- *External Validity*
 - Does sample represent population?
 - Do instruments measure phenomena of interest?
 - Can the findings be generalized?

Bias

- Any systematic error in an epidemiologic study that results in an incorrect estimate of the association between exposure and risk of disease.
- Types of bias:
 - *Selection Bias*
 - *Information Bias*

Selection Bias

The error arising from differentials in identifying the study populations (disease or exposure status)

Occurs when inclusion in the “disease” group depends on the “exposure” and vice versa.

Selection Bias is likely to occur in Case-control studies and Retrospective follow-up studies.

It is unlikely to occur in Prospective follow-up studies because exposure is ascertained before disease occurrence.

Common sources of selection bias are:

Using referred or hospitalized cases which depended on prior knowledge of the exposure-disease relation

Non-response or refusal in case-control studies if related to exposure status

Information Bias

Also known as observation bias.

Results from systematic differential in the way data is obtained about exposure or disease.

Recall bias

Interviewer bias

Loss to follow-up

Misclassification

(when random, it dilutes the association)

Bias can mask a true association or produce a spurious association

Elimination of bias

Opportunity:

In the study design

In data collection

It is *not always possible* to eliminate selection or information bias; however rigorous procedures must be adopted to *reduce such bias*. Potential sources of bias should be discussed in the results.

Confounding

- Is bias in the estimation of the effect of exposure on disease occurrence, due to a *lack of comparability (lack of exchangeability)* between exposed and unexposed populations; thus, disease risks would be different even if the exposure were absent in both populations.
- A third factor which is related to both exposure and outcome, and which accounts for some/all of the observed relationship between the two.

Confounding

- Confounder not a result of the exposure
 - e.g., association between child's birth rank (exposure) and Down syndrome (outcome); mother's age a confounder?
 - e.g., association between mother's age (exposure) and Down syndrome (outcome); birth rank a confounder?

Confounder

- Is a factor that distorts the true relationship between an exposure and the disease outcome on account of its being associated with both the exposure as well as the disease.
- This distortion (over/underestimation) of the true relation between exposure and disease can occur only if this factor is unequally distributed between the exposed and unexposed groups.

General Rule

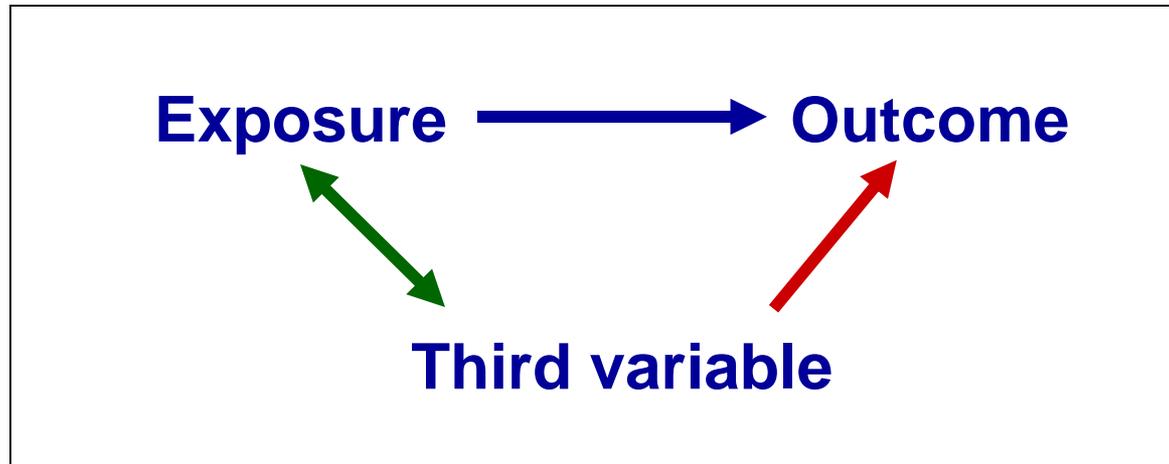
To be a confounding factor:

A variable must be:

- *Causally associated with the outcome “disease”*
- *Causally or non-causally associated with “exposure”*
- *Not an intermediate factor in the “disease-exposure” causal pathway*

Confounding

To be a confounding factor, two conditions must be met:



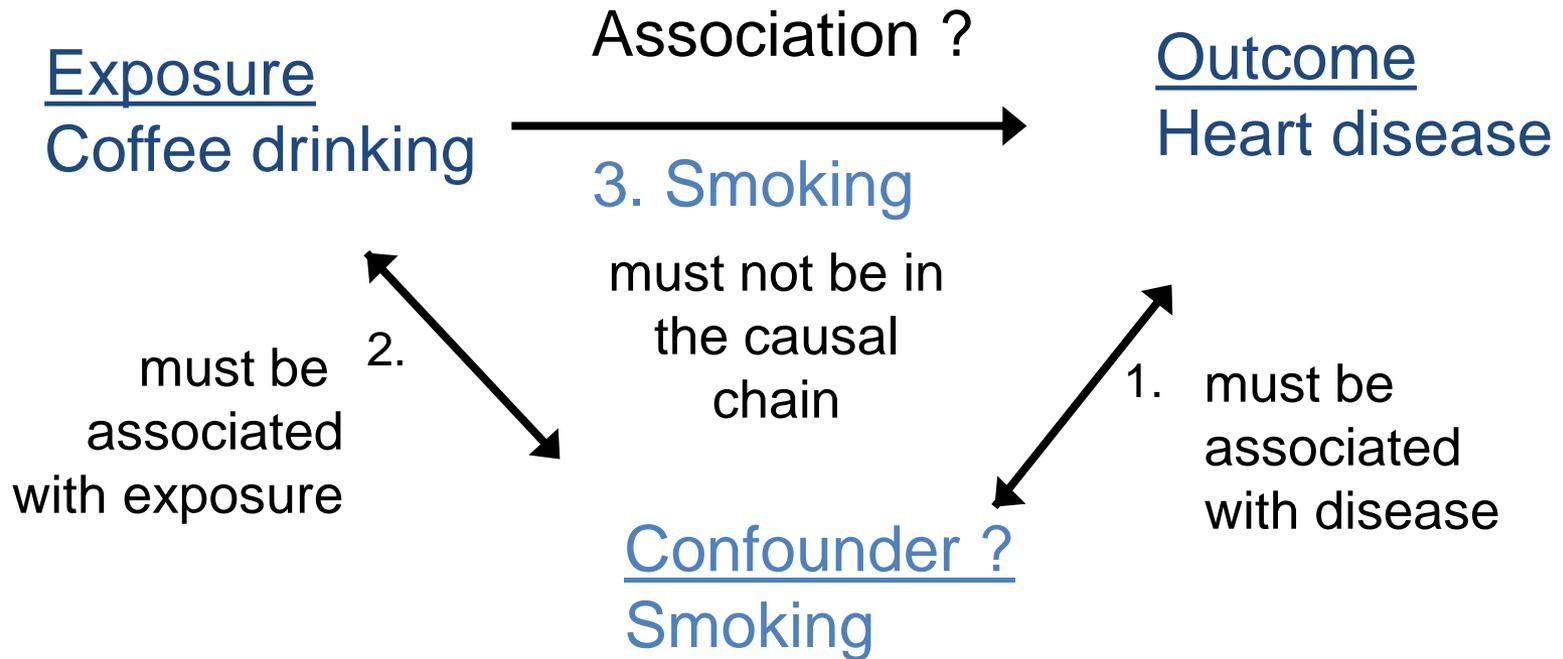
Be associated with exposure

- without being the consequence of exposure

Be associated with outcome

- independently of exposure (not an intermediary)

Confounding



Confounding example: Coffee and myocardial infarction (MI)

	<i>MI</i>	<i>NO MI</i>
<i>COFFEE</i>	<i>90</i>	<i>60</i>
<i>NO COFFEE</i>	<i>60</i>	<i>90</i>

Crude OR= $90 \times 90 / 60 \times 60 = 2.25$

Positive association between coffee and MI

Is this confounded by smoking ?

Confounding: Coffee and myocardial infarction (MI)

	<i>MI</i>	<i>NO MI</i>
<i>SMOKING</i>	<i>100</i>	<i>50</i>
<i>NOT SMOKING</i>	<i>50</i>	<i>100</i>

$$OR = 100 \times 100 / 50 \times 50 = 4$$

Positive association between smoking and MI

Therefore – smoking is associated with disease

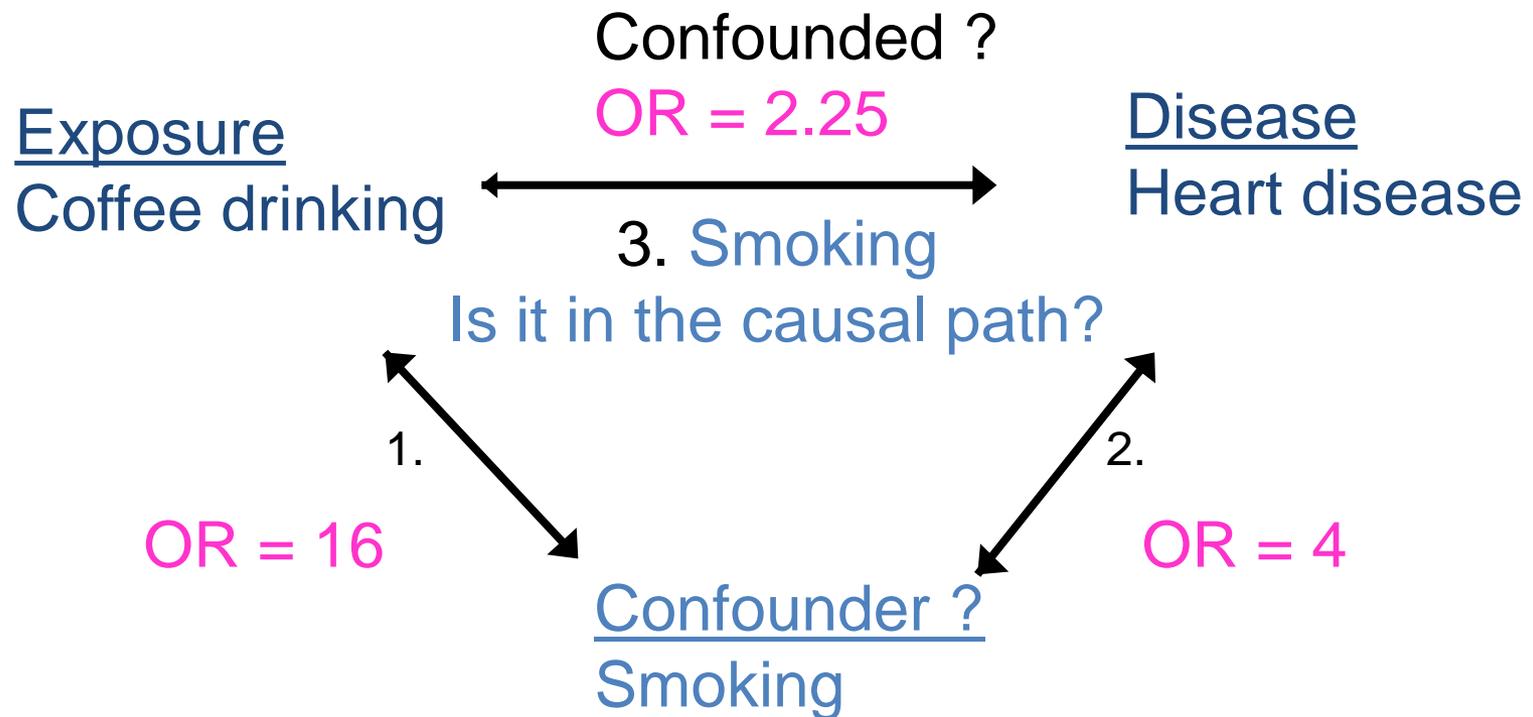
Confounding: Coffee and myocardial infarction (MI)

	<i>SMOKER</i>	<i>NONSMOKER</i>
<i>COFFEE</i>	120	30
<i>NO COFFEE</i>	30	120

$$OR = 120 \times 120 / 30 \times 30 = 16$$

Positive association between coffee and smoking therefore – smoking is associated with exposure

Confounding: Coffee and myocardial infarction (MI) - 4



*Confounding: Coffee and myocardial infarction (MI)
Exploring through stratification*

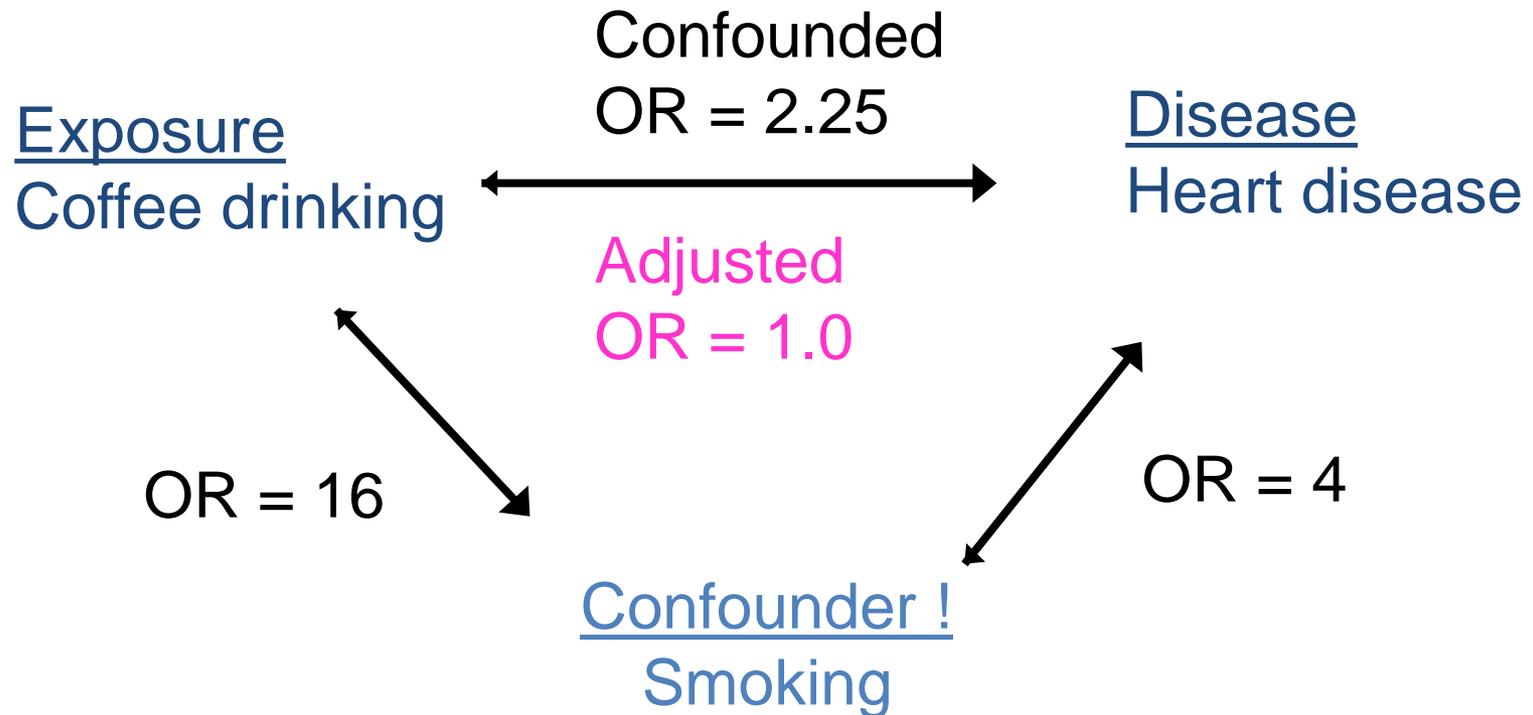
	<i>Smokers</i>		<i>Nonsmokers</i>	
	<i>MI</i>	<i>No MI</i>	<i>MI</i>	<i>No MI</i>
<i>Coffee</i>	<i>80</i>	<i>40</i>	<i>10</i>	<i>20</i>
<i>No coffee</i>	<i>20</i>	<i>10</i>	<i>40</i>	<i>80</i>

$$OR = 80 \times 10 / 20 \times 40 = 1.0$$

$$OR = 10 \times 80 / 40 \times 20 = 1.0$$

There is no effect of coffee, when smoking is held constant therefore - smoking is acting as a confounder

Confounding: Coffee and myocardial infarction (MI) - 6



Managing a confounder

Which variables to include?

Depends on knowledge and experience from previous studies and logic.

How do we know about it?

We do the analysis for the crude estimate.

We then do the analysis with control of the suspected factor.

If we observe a difference in the estimates, then that variable is a confounder.

If the information is not there, then we can not identify or control the confounder.

Steps to explore confounding

1. Is there an association?
2. If so, is it due to confounding?
 - NO → Likely causal
 - YES → Not causal
3. Is the association equally strong in strata formed on the basis of a third variable?
 - NO → Interaction (effect modification) is present
 - YES → Interaction (effect modification) is NOT present

Effect Modification will be dealt with separately

Control of Confounding

Control in the Design:

Randomization

Restriction

Matching

Control in the Analysis:

Stratified Analysis

Multivariate Analysis

Randomization

- *The procedure of choice in Intervention (Experimental) studies through random allocation of subjects to various study groups.*
- *It's unique strength is control of confounding.*
- *If study sample sufficiently large, randomization virtually insures elimination of known and unsuspected confounding factors.*
- *Since unknown confounders can not be controlled by analysis, only randomization (with sufficient study size) can eliminate them*

Restriction

No confounding if no variation of the variable in either exposure or disease categories.

Restrict admission to one category of the variable, e.g. males only, age (within a narrow bracket) etc.

Limitations:

Reduces sampling frame

Residual confounding (if bracket not narrow enough)

Does not allow evaluation of the various categories of the variable, e.g. females and age outside bracket.

Matching

- *Includes elements in the design and analysis.*
- *Primarily used in case-control studies*
- *Example: In the study of MI and Exercise, controls were matched with cases for age, gender and level of smoking.*
- *Matching used to be very appealing, but usually cumbersome, time consuming and expensive.*
- *Alternative analysis techniques overshadowed matching.*
- *If matching is done matched analysis must be carried.*

Matching (cont.)

- *The effect of the matched factor on the risk of disease can not be evaluated (similar to restriction)*
- *Best advantage is matching for variables that are complex and difficult to quantify. Examples:*
 - *Siblings for factors usually strongly correlated in family members, such as: early environmental exposures, genetic factors, dietary habits, SES and health care facilities.*
 - *Neighborhood is a surrogate for environmental exposure and Socio-economic status*

Matched pair analysis

Presentation of data of a matched pair case-control study

		<i>Controls</i>		
		<i>Exposed</i>	<i>Non-exposed</i>	
<i>Cases</i>	<i>Exposed</i>	<i>a</i>	<i>b</i>	<i>a+b</i>
	<i>Non-exposed</i>	<i>c</i>	<i>d</i>	<i>c+d</i>
	<i>Total</i>	<i>a+c</i>	<i>b+d</i>	

(a) and (d) are called concordant pairs

(b) and (c) are called discordant pairs

All the information is in the discordant pairs

Matched pair analysis

Presentation of data from a matched pair case-control study

Cases	Controls	
	Exposed	Non-exposed
Exposed	39	113
Non-exposed	15	150

Data from study of exogenous estrogens and endometrial Carcinoma. Source DC Smith et al, NEJM 293: 1164, 1975

Analysis and testing of matched pair data

Estimation:

$$\text{Odds Ratio} = \mathbf{OR} = \frac{b}{c}$$

Testing:

$$\text{McNemar test: } \chi_{(1)}^2 = \frac{(b-c)^2}{(b+c)}$$

$$\mathbf{OR} = 113/15 = 7.5$$

$$\mathbf{Ch^2} = (113-15)^2 / (113+15) = 75.03$$

CI can be computed using the variance or the test-based procedures

Stratified Analysis

Estimation Using the standard tables of data presentation:

Only, this time stratified by confounding variable:

For cohort with count data or case-control data:

$$RR_{MH} = \frac{\sum a(c+d) / T}{\sum c(a+b) / T} \quad \text{and} \quad OR_{MH} = \frac{\sum ad / T}{\sum bc / T}$$

Multivariate Analysis

You should not worry about these now:

Your biostatistics course will cover it.

The computer will do all the work for you if you know what to request from it.

Multivariate Analysis

Multiple Linear Regression

Used principally with continuous (measured) outcome variables

Extension of the simple linear expression:

$$Y = a + bX \quad \text{to:}$$

$$Y = a + b_1X_1 + b_2X_2 + \dots + b_nX_n$$

Where:

Y = mean of dependent (outcome) variable

X = set of independent (predictor) variables

b = coefficient for each independent variable

Assumptions:

A linear function of the variables in the model

Interpretation:

*b = Estimated mean change in Y for each unit change in X,
while controlling for the confounding effect of others*

Xs could be continuous, categorical or dichotomous variables

Multivariate Analysis

Multiple Logistic Regression

Used principally with binary outcome variables

Outcome (dependent) variable is the natural logarithm (ln) of the odds of disease (logit)

Ln {Y/(1-Y)} is a linear function of predictor variables

$$\ln \left[\frac{y}{1-y} \right] = a + b_1 X_1 + \dots + b_n X_n$$

Summary

- *Error and bias* can lead to misclassification
- *Confounding & effect modification* can also bias association
- *Misclassification* may result in a larger (or smaller) association being calculated
- *Association* may not be causal, but due to misclassification – or -
- *Lack of association* may mask true causal relationship
- Statistical tests do not evaluate bias; only chance

Errors in Epidemiological Studies

- **Random Error**
 - **Sample Size Calculations**

Random Error

- **Divergence, due to chance alone, of an observation on a sample from the true population value, leading to lack of precision in the measurement of an association**
- **Sources of Random Error**
 - **Sampling error**
 - **Biological variation**
 - **Measurement error**

Random error

- A measurement error whose value varies randomly in measuring the same value of quantity in same conditions. Random error can't be removed with calibration. It has a specific distribution with an average value and the distribution deviation can be approximated. When the deviation is known, the range of the random error can be forecasted with statistical methods.

Random errors

- An error that varies between successive measurements
- Equally likely to be positive or negative
- Always present in an experiment
- Presence obvious from distribution of values obtained
- Can be minimised by performing multiple measurements of the same quantity or by measuring one quantity as function of second quantity and performing a straight line fit of the data
- Sometimes referred to as reading errors

Random Error

- Results from variability in the data, sampling
 - *E.g.* measuring height with measuring tape: 1 measurement may be off, but multiple measurements will give you a better estimate of height
- Relates to precision
- We use confidence intervals to express the degree of uncertainty/random error associated with a point estimate (e.g. a RR or OR)
 - Measure of precision

Sample Size Calculations

Variable to consider

- Required level of statistical significance of the expected result
- Acceptable chance of missing the real effect
- Magnitude of the effect under investigation
- Amount of disease in the population
- Relative sizes of the groups being compared

Causal Relation

Causation

Causation is any cause that produces an effect.

This means that when something happens (cause) something else will also always happen(effect).

An example: When you run you burn calories.

As you can see with the example our cause is running while burning calories is our effect. This is something that is always, because that's how the human body works.

Correlation

Correlation measures the relationship between two things.

Positive correlations happen when one thing goes up, and another thing goes up as well.

An example: When the demand for a product is high, the price may go up. As you can see, because the demand is high the price may be high.

Negative correlations occur when the opposite happens. When one thing goes up, and another goes down.

A correlation tells us that two variables are related, but we cannot say anything about whether one caused the other.

Correlation

Correlations happen when:

A causes B

B causes A

A and B are consequences of a common cause, but do not cause each other

There is no connection between A and B, the correlation is coincidental

Causation and Correlation

Causation and correlation can happen at the same time. But having a correlation does not always mean you have a causation.

A good example of this:

There is a positive correlation between the number of firemen fighting a fire and the size of the fire. This means the more people at the fire, tends to reflect how big the fire is. However, this doesn't mean that bringing more firemen will cause the size of the fire to increase.

Correlation or Causation?

As people's happiness level increases, so does their helpfulness.

This would be a correlation.

Just because someone is happy does not always mean that they will become more helpful. This just usually tends to be the case.

Was it Clear Enough !

